DOCUMENT RESUME

ED 206 059 EA 013 781

AUTHOR	Barker, Pierce: And Others
TITLE	A Study of Alternatives in American Education, Vol. V: Diversity in the Classroom.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	Rand-R-2170/5-NIE
PUB DATE	Feb 81
CONTRACT	B2C-5326
NOTE	207p.: For related documents, see EA 013 780, ED166 825, and ED 163 591.
AVAILABLE FROM	The Rand Corporation, 1700 Main St., Santa Monica, CA 90406 ($7.00).
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	*Demonstration Programs: Educational Research: *Educational Vouchers: Elementary Secondary Education: *Free Choice Transfer Programs: *Nontraditional Education: Public Schools
IDENTIFIERS	Alum Rock Union School District CA: *Educational Diversity: School Choice

ABSTRACT

An examination of the variation among educational programs in the Alum Rock Union Elementary School District, San Jose (California), during the fourth year of its educational voucher demonstration project addresses the issue of whether alternative systems can generate significant diversity in public school education. The research questions (How do individual classrooms differ? What variables primarily distinguish program types? Do voucher demonstration classrooms differ, on the whole, from nonvoucher classrooms?) and policy issues were pursued primarily through multidimensional scaling of structured observation data collected in third- and fifth-grade demonstration and control classrooms. On the measures of diversity employed, voucher classrooms were found to be remarkably similar to each other. The reasons for the diversity from nonvoucher classrooms cannot be easily identified, due to incomplete sampling procedures. It is also difficult to ascertain whether differentiating features could have arisen without the voucher demonstration. The two or three classrooms in the sample that were strikingly different in organization and operation are probably the results of unexplained teacher preferences. The report concludes that the voucher demonstration classrooms seem to be characterizable as diverse and that the measures employed detected limited variation among classrooms. (Author/WD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206059

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A STUDY OF ALTERNATIVES IN AMERICAN EDUCATION
VOL. VI: RESEARCH IN THE CLASSROOM

EA 013 781

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

T. Cockerell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The work upon which this publication is based was performed pursuant to Contract B2C-5326 with the National Institute of Education, Department of Health, Education, and Welfare.

The Rand Publications Series: The Report is the principal publication documenting and transmitting Rand's major research findings and final research results. The Rand Note reports other outputs of sponsored research for general distribution. Publications of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation

A STUDY OF ALTERNATIVES IN AMERICAN EDUCATION, VOL. V: DIVERSITY IN THE CLASSROOM

PREPARED FOR THE NATIONAL INSTITUTE OF EDUCATION

PIERCE BARKER, TORA BIKSON, JACKIE KIMBROUGH,
WITH THE ASSISTANCE OF CAROL FROST

R-2170/5-NIE
FEBRUARY 1981

$7.00

Rand
SANTA MONICA, CA. 90406

PREFACE

This report is the fifth volume of a series documenting a study of alternative schools in American education, sponsored by the National Institute of Education under Contract B2C-5326. There are six other volumes in the series, all published or forthcoming under the general title, A Study of Alternatives in American Education:

Vol. I: District Policies and the Implementation of Change, by G. Bass, R-2170/1-NIE.

Vol. II: The Role of the Principal, by M. Thomas, R-2170/2-NIE.

Vol. III: Teachers' Responses to Alternatives, by R. Rasmussen, R-2170/3-NIE.

Vol. IV: Family Choice in Schooling, by R. G. Bridge and J. Blackman, R-2170/4-NIE.

Vol. VI: Student Outcomes in Alum Rock, 1974-1976, by F. J. Capell, R-2170/6-NIE.

Vol. VII: Summary and Policy Implications, by W. Furry and D. Weiler, with Joyce Peterson, R-2170/7-NIE.

STUDY BACKGROUND

This study had its origins in 1972. In April of that year, the Office of Economic Opportunity (OEO) funded an education voucher demonstration in Alum Rock, California, and awarded a study and evaluation contract to The Rand Corporation. Voucher systems require that funds for education be distributed directly to families in the form of certificates, which families can then use to purchase education at schools of their choice.[1] The government wished to test a voucher model that included competing public and private schools, with complex regulations

[1] Findings for the first year of the voucher demonstration (1972-73) are reported in Daniel Weiler et al., A Public School Voucher Demonstration: The First Year at Alum Rock, The Rand Corporation, R-1495-NIE, June 1974, 4 vols. Alum Rock is an independent elementary school district in San Jose, California.

designed to protect and advance the interests of disadvantaged families.[2] But the OEO agreement with Alum Rock did not require immediate implementation of this model. In lieu of private schools participating in the demonstration, Alum Rock was to encourage parent choice and stimulate competition between schools--two key objectives of the voucher plan--by creating multiple programs within the public schools. Parents would be informed about their options and encouraged to select the programs they preferred for their children. Alum Rock and OEO agreed that this "public schools only" model was to be a "transition" toward a more complete voucher demonstration, and OEO continued to seek additional demonstration sites for a more extensive test of the voucher idea. The demonstration began in September 1972 with six schools, organized as twenty-two "minischools" offering a variety of educational approaches.

By the end of the second year of the demonstration--spring 1974--sponsorship of the voucher program had been assumed by the National Institute of Education. The transition to a full-scale model in Alum Rock had not taken place, and no new sites had joined the demonstration. Rand and NIE agreed, however, that while a more complete voucher test might still be arranged in Alum Rock or elsewhere, the existing demonstration was of interest in its own right: Thirteen public schools were offering forty-five program options to parents.[3] In effect, Alum Rock was testing a variant of an innovation that a number of observers had argued could improve the quality of public education--alternative schools.

It was agreed that while the main study would continue to concentrate on Alum Rock in 1974-75, a small side study would be undertaken to explore the nature of the alternative schools movement in other

---

[2]The "regulated compensatory" voucher model was originally proposed in a 1970 study commissioned by OEO. See Center for the Study of Public Policy, Education Vouchers: A Report on Financing Elementary Education by Grants to Parents, Cambridge, Mass., December 1970.

[3]There were at one time more than fifty mini-schools available to participating parents, in fourteen demonstration schools. Ten Alum Rock schools never joined the demonstration.

districts. This study identified a number of areas where further analysis might yield a better understanding of the issues associated with implementing alternative schools. Many of these issues had already surfaced in Alum Rock.

By the fourth year of the demonstration (1975-76), prospects for creating a more comprehensive test of the voucher model had diminished appreciably, while the work that had already been accomplished in Alum Rock constituted a useful base for a modest comparative study of alternative schools. Accordingly, some project resources were shifted in that year toward the study of three new sites where alternative schools were being tried: Cincinnati, Ohio; Eugene, Oregon; and Minneapolis, Minnesota.[4] Data collection from these sites and Alum Rock was completed in 1976-77.

ALTERNATIVE SCHOOLS

Alternative schools or educational programs--variously defined--can now be found in perhaps one out of every four school districts in the country.[5] These schools and programs serve a number of different client groups, offering some form of teaching style and method or curriculum content differing in important respects from the mainstream of educational programs in those districts. They have been created in response to a variety of social and political pressures, and are usually designed to meet some or all of the following objectives:

- Social Equity. Extending to all parents the right to choose among educational alternatives that they consider best suited for their children, and/or reinforcing area- or district-wide desegregation plans by providing "magnet" programs.

---

[4] Criteria and methods for site selection are discussed in Chapter 1 of Vol. I in this series: District Policies and the Implementation of Change, by G. Bass.

[5] National School Boards Association, Alternative Schools, Research Report 1976-3, Evanston, Illinois, 1976, p. 5.

- Accountability. Creating schools that are more directly responsive to parent and student desires and needs, and more visible and open regarding their educational operations.
- Incentives to Innovate. Proving expanded opportunities for teachers and administrators to offer new and different educational programs, where rewards for successful innovation are tied in part to extrinsic evaluations of success (parent and student demand) rather than (exclusively) to intrinsic criteria (administrator and colleague approval).
- Diversity. Introducing program variety on the assumption that a uniform approach to education may be inefficient where the student clientele is socially and ethnically diverse, with a range of skills and interests that should be matched to appropriately varied educational opportunities.
- Constituency Satisfaction. Increasing student, parent, teacher, and community satisfaction with the education system, both as a desirable social goal in itself and as the means to other desirable objectives: increased social stability in the schools, greater parent support of school activities, and more community willingness to provide financial support for education.
- Improved Student Outcomes. Improving student cognitive and noncognitive growth through better matching of students to programs, through improved teacher and administrator incentives to innovate, and through greater system accountability.

For the purposes of this study, an alternative school or program is defined as having at least three essential characteristics:

- It is an educational program that is distinctly different in some way from the majority of programs in that district.
- It is available to students on a voluntary basis.
- It is a full-time educational program.

The study is confined to alternatives that meet this definition.[6]

STUDY GOALS AND CONSTRAINTS

As noted, this study had its origins in Rand's evaluation of the voucher demonstration in Alum Rock. In that district, rapid and complex changes in organization and procedure raised many questions about the problems that a district might encounter in attempting to implement a system of alternative schools. These questions became the main focal points of the research, as hypotheses generated on the basis of Rand's study of Alum Rock were tested against the experiences of other districts that have tried some version of an alternative schools program. We asked: What district strategies are most likely to lead to the successful implementation of alternatives under different circumstances? What are the effects of alternative programs on teacher behavior, and how do teachers influence the outcomes of such programs? What role does--or should--the school principal play? How do parents react--do they understand their choices? If so, how do they exercise their options? Is real diversity possible within the public schools, with the many internal and external pressures to conform to a common program?

The seven reports in this study address these and related questions. The study is aimed at practitioners and community groups who may be contemplating the initiation of alternative schools, and at state and federal policymakers who may be asked to support alternatives and would like to be aware of the obstacles and opportunities that this innovation can create.

The study draws no conclusions about the relative desirability of alternatives; this is a value judgment that citizens and professionals

---

[6] In practice, we were obliged to select districts for the study on the basis of claimed and apparent program distinctiveness; the extent of actual distinctiveness was then treated as an issue to be explored in the course of the analyses. A discussion of the recent history of the alternatives movement will be found in Chapter 1 of Vol. I of this series.

must make on the basis of what they want from their schools. Thus, although individual authors have tried to make their own normative positions clear, nothing in these reports should be interpreted as representing a Rand position on alternatives as an innovation in public education. Nor is this an evaluation: We have studied four districts to gain insight into common (and uncommon) problems; nothing in these reports should be construed as presenting evaluative judgments about the advisability or wisdom of any district policy. Finally, this study is not an assessment of the educational impact of alternatives on students. With the exception of Alum Rock--a unique case--the study did not have access to the longitudinal student outcome data that would have to be analyzed to make such an assessment.

The study utilizes a small purposive sample that was designed to select sites where an important effort had been made to implement alternatives. For reasons explained in a number of the reports, we believe that it is legitimate to assume that many of the study's findings will have wide applicability. Nevertheless, the reader should bear in mind the limited and selective nature of this sample when considering the generalizability of study results.

ORGANIZATION OF THE STUDY

The study findings are organized as a series of reports on the issues of implementing alternatives from the perspectives of major participants: district administrators, principals, teachers, and parents. A fifth report focuses on the extent of program diversity achieved in the Alum Rock mini-school system, a sixth report presents an analysis of student outcomes in Alum Rock, and a final report provides an overview of the entire study. The reports are related, but each is also designed to be read as an independent study. The reader who completes the entire series will therefore notice some redundancy: Each report begins with a similar discussion of study methods and study site settings. There is also some inevitable overlap in the discussion of key issues, since a report that deals with any part of the educational system must to some extent discuss other aspects of the system as well.

Thus, for example, the report that focuses on the role of the principal can hardly avoid discussing the views of teachers, and vice versa. In the interest of writing reports that are independent research documents as well as part of a general study topic, we have made no attempt to eliminate these redundancies.

Since 1972, this study has produced 39 informal Working Notes for client (OEO and NIE) use, mostly on selected aspects of the Alum Rock demonstration. These documents, together with project Administrative Reports and original materials (documentary materials, surveys, and field notes) form an extensive primary and secondary data base, which has been drawn on as needed by the authors of study reports, largely without specific citation. Where it is appropriate to call the reader's attention to a particular source of evidence in the informal secondary materials, the latter are cited as unpublished papers. These and related unpublished materials are available from the National Institute of Education.

The introductory chapter of each report in the series provides further details about the particular focus, methods, and limitations of that report. The work reported in this volume is specifically addressed to the empirical determination of education alternatives. It investigates the question of whether or not Alum Rock mini-schools succeeded in meeting the first defining characteristic of alternative schooling, that of providing educational programs distinctly different in some way from the majority of programs in the district. Such a study is expected to serve at least three purposes. First, it provides a methodological paradigm for constructing an a posteriori typology of educational programs from observational and teacher self-report data. Second, the data so obtained illuminate questions about the extent to which limited voucher procedures of the sort introduced in the Alum Rock district are likely to produce a real rather than a nominal diversity of educational options for parent choice. Finally, in providing an account of the degree to which really distinct alternatives were implemented in that district, the study establishes a basis for interpreting and assessing the effects of the demonstration

on student outcomes (Vol. VI). Understanding discussions of methodology and analysis in the text requires some familiarity with statistics and multivariate analysis.

Daniel Weiler
Study Director

SUMMARY

As part of its commitment to improving public education, the federal government has subsidized alternative schools and programs across the nation. A major objective of these schools and programs is educational diversity. By offering program content or using teaching methods that depart significantly from its traditional programs, a district can, ideally, meet the needs of a socially and ethnically diverse student population. An important question for educators and federal policymakers is whether that objective has been or can be met, whether programs are actually or only nominally alternatives. To answer that question, they need systematic means of measuring the actual diversity achieved in districts that offer such programs. The Rand research project reported here developed and tested instruments for making these measurements, and applied them to a multiple-option school district.

The project was part of a larger Rand study of alternative schools, sponsored by the National Institute of Education. The purpose of the project was to discover whether a voucher system could generate real diversity of options in public education. (A voucher system allows parents to choose among educational alternatives by providing them with certificates representing the public education funds available to their children. Parents can use these certificates to "purchase" education at the schools of their choice.) To answer that important policy question, we undertook a description of the options that had developed during the demonstration. A basic assumption of our study was that nominal difference does not guarantee diversity. We attempted to discover, inductively, the characteristics of learning environments that indicate diversity--that is, significantly different educational alternatives.

Data for this study came from the Alum Rock Union Elementary School District (San Jose, California) during 1975-76, the fourth year of an Educational Voucher Demonstration. The fourth year was chosen in order to allow time for alternative programs to be fully established. Forty voucher and 34 non-voucher classrooms constituted the analytic sample

for the study. The voucher classrooms represented 19 mini-schools within 8 elementary schools. (Mini-schools are instructional units that provide educational options and compete for students and resources within and across schools.) The comparison group of 34 non-voucher classrooms represented nine elementary schools.

In order to determine whether the educational programs of the mini-schools were empirically distinguishable, and whether the classrooms within the programs reflected the pattern of diversity expected to emerge in the demonstration, we developed variables from three sources: the mini-schools' self-reports, educational literature, and our pilot studies in Alum Rock. We examined these variables to see which could be reliably observed and which were useful for discriminating among Alum Rock classrooms. The variables ultimately selected were used to construct data-collection instruments for trained observers to complete and questionnaires to be filled out by the teachers whose classes were observed. We relied primarily on observation rather than self-reports because: programs differ in ways not apparent to participants; nominally distinct programs may be similar and nominally similar programs, distinct; and participants may not be able to judge how fully implemented a program is.

We used the datasets generated by the observations and questionnaires to construct and interpret a three-dimensional, spatial model of the diversity among educational alternatives. In the model, proximity among units indicates similarity among classrooms and distance indicates dissimilarity.

The model reveals that for Alum Rock, at least, the voucher system has not generated educational alternatives that are truly diverse. With the exception of three "outlying" classrooms, the rest of the voucher and all of the non-voucher classrooms cluster around the center of the model. In spread and orientation, the voucher classrooms are strikingly similar to the non-voucher classrooms. Despite their intended differences in content or method, the alternative programs are not measurably different from the standard program represented by the non-voucher classrooms.

We would not conclude from these findings that alternative programs cannot generate educational diversity. The study has a number of limitations that would make such a generalization irresponsible. However, the results do suggest that our approach is a valid means of detecting and portraying variation among educational alternatives. The three outlying classrooms were empirically different, in terms of the variables measured, from the others observed. Further, the self-reports of the teachers, outliers and the rest, were reasonably congruent with their teaching practices as embodied in our observations. Thus, the model does capture actual diversity.

The voucher demonstration encouraged new approaches and should have provided an excellent opportunity for teachers to try out different methods and contents. Our data cannot explain why only three voucher classrooms were strikingly different in organization and operation. The most reasonable speculation is that these outlying classrooms reflect educational preferences of the individual teachers, preferences not shared by their other colleagues. It may be that these outliers' predemonstration classrooms would have exhibited the same characteristics that their voucher classrooms did. On the other hand, decentralized decisionmaking introduced by the demonstration may have allowed them to put their preferences into action. At any rate, the voucher system in Alum Rock permitted but did not promote an observable diversity of educational alternatives.

ACKNOWLEDGMENTS

We wish to acknowledge with gratitude a number of individuals who assisted in the completion of this research. Of invaluable help were Dr. William Jefferds, Superintendent of the Alum Rock Unified School District, and the principals and minischool coordinators whose cooperation was essential at every stage of the work. Special thanks are owed to Dr. Thomas Bikson (The Rand Corporation) for his development of data management procedures and participation in observer training, as well as to Jose Hernandez (ARUSD Evaluation Department) for his on-site supervision of data collection activities. Recognition is also due the 17 substitute teachers who served as research assistants, conducting observation sessions with competence, patience, and good spirit. We are especially indebted to Dr. Courtenay Cazden (Harvard University) and Dr. Eric Holman (University of California, Los Angeles) for their careful critical reviews and comments on the first draft of this report. Finally, we warmly thank the Alum Rock elementary school teachers whose classrooms constituted the subject of this study.

CONTENTS

PREFACE iii

SUMMARY xi

ACKNOWLEDGMENTS xv

Chapter

1. INTRODUCTION 1
   - Research Questions 1
   - District Context 2
   - Policy Considerations 4
   - Procedures 5
   - Analyses 7
   - Limitations of the Study 8

2. CONCEPTUAL BACKGROUND 11
   - Approach to the Study of Classroom Diversity 11
   - Refinement and Operationalization of Variables 22

3. PROCEDURES 32
   - Instruments 32
   - The Sampling Universe 39
   - The Sampling Plan 40
   - Data Collection 42
   - Preliminary Meetings with District Personnel 44
   - Pilot Study 45

4. ANALYSIS 59
   - Description and Definition of Diversity Variables 61
   - Construction of Descriptive Spaces 64
   - Distances 68
   - Analytic Samples 73
   - Basic Scaling Results 74

5. INTERPRETATION 86
   - Nonvoucher Classrooms 97
   - Summary of Analysis by Data Source 101

6. DISCUSSION 104
   - Diversity Reconsidered 106
   - Future Research 107

Appendix

A. INSTRUMENTS 109

B. OPERATIONAL DEFINITIONS OF TERMS IN THE OBSERVATION INSTRUMENT 123

C. ANALYTIC SAMPLE DESIGN FOR THE OBSERVATIONAL STUDY OF DIVERSITY 146
D. DERIVED SCALES FROM CLASSROOM OBSERVATION AND TEACHER QUESTIONNAIRE 173
E. PROGRAMS EMPLOYED IN ANALYSES 179

BIBLIOGRAPHY 185

TABLES

2.1. Pilot Factor Analysis 30

3.1. Classroom Observer Assignment 47

3.2. Interobserver Agreement for Pilot Study 49

3.3. Design Chart for Observed Voucher Classrooms 55

4.1. Coordinates of Defining Classrooms In 2- and 3-Space 77

5.1. Correlation of Scale Values and Projections 91

C.1. Factors Explicit in Hypothesized Generative Model of Observations 148

C.2. Full Anova Model With All Factors Minimally Confounded .. 152

C.3. Variance Attributable to Artifact or Unreliability 154

C.4. Schematic Illustration of Successful Hierarchical Partitioning of Variation Between Classrooms 155

C.5. Full Structural Model 158

C.6. Anova Model for Preliminary Reliability Assessment 162

C.7. Structural Model for Reliability Design 163

C.8. Example of Partial Confounding in C/M/S Hierarchy Spring 1975 Data 167

C.9. Example of Loss of Degrees of Freedom From Confounding in Sample Spring 1975 Data 168

C.10. Symbolic Layout of the Effects of Introducing a Mini-School Typology that Crosses Schools 171

FIGURES

4.1. V Classrooms in the 3-Space Configuration 76

4.2. V Classrooms and Mini-School Connections: Dimension I x Dimension II 82

4.3. V Classrooms and Mini-School Connections: Dimension I x Dimension III 83

4.4. V Classrooms and Mini-School Connections: Dimension II x Dimension III 84

5.1. V Classrooms in the 3-Space Configuration with Interpretive Vectors 90

5.2. V and NV Classrooms Scaled in Same Space: I x II 98

5.3. V and NV Classrooms Scaled in Same Space: I x III 99

5.4. V and NV Classrooms Scaled in Same Space: II x III 100

Chapter 1

INTRODUCTION

RESEARCH QUESTIONS

This study examines the variation among educational programs in the Alum Rock Union Elementary School District in San Jose, California, during the fourth year (1975-76) of the Educational Voucher Demonstration, which operated from 1972 to 1977. The investigation explores the nature and extent of educational program diversity generated by the demonstration on the classroom level, and contributes to research on educational options in three respects. First, it responds to the need cited by Brandt (1973) and Stallings (1974) for accurate representation of educational inputs through its description of specific educational programs. Second, the study analyzes the accuracy of the district's descriptions of the educational options it offered by seeing how closely they matched the options as they were implemented in the classroom. Third, the research describes variables and classes of variables that account for observable differences among the classrooms.

To address the global issue of whether alternative systems can generate significant diversity in public school education, the research sought to distinguish the kinds of educational alternatives developed by the voucher demonstration. The district offered more than 50 educational programs during the fourth year of the demonstration, making a *prima facie* case that it achieved educational diversity; but even widespread individual differences among classrooms do not guarantee a significant diversity of options. Rather, educational diversity requires not only that classrooms differ substantially along important dimensions of the learning environment, but also that different features be systematically integrated to provide a set of coherent educational alternatives.

Hence, this analysis attempts to determine whether Alum Rock established empirically distinguishable learning environments, and explores the dimensions that give structure to program options. Drawing primarily on observation data, the study seeks to answer the following questions:

1. How do individual classrooms differ? Can these differences be summarized by locating the classrooms in a multidimensional space?
2. What variables primarily distinguish program types? That is, what dimensions of educational practice interpret the spatial model of classroom differences?
3. Do empirically obtained relationships among classrooms exhibit the spatial structure to be expected on the basis of school and/or mini-school organization?
4. Are there ways in which voucher demonstration classrooms differ, on the whole, from nonvoucher classrooms?

DISTRICT CONTEXT

The Alum Rock Union Elementary School District (ARUESD) is one of 11 school districts serving the city of San Jose, California, whose 1970 population was about 445,000.[1] San Jose is one of the fastest growing cities in the state. The ARUESD serves a largely minority, working-class community of some 46,000 in the eastern section of San Jose. Median family income in Alum Rock in 1970 was approximately $10,000, with 10 percent of the families below the poverty level, in contrast to 5.6 percent for San Jose as a whole. Alum Rock's low income is also reflected in school enrollment statistics: 35 percent of its students received Aid to Families with Dependent Children (AFDC) in 1976-77, the highest percentage among all districts in Santa Clara County, and about 72 percent participated in the subsidized school lunch program.[2]

The school district, which serves kindergarten through eighth grade, had 18 elementary schools and 5 middle schools in 1975-76, the year of our study. In the fall of 1975, almost 70 percent of its

---

[1] Population and income figures were derived from the 1970 Census of Population and Housing, San Jose SMSA.

[2] School and student characteristics were obtained from the ARUESD Research and Evaluation Office.

14,000 students were from minority groups, with 57 percent of the students Hispanic, 12 percent black, and 2-1/2 percent of Asian, Pacific Island, and native American extraction. In contrast, the majority (73 percent) of the district's 588 regular teachers were white, 12 percent Hispanic, 8 percent black, and 6 percent of Asian and native American origin.

The district budget was more than $21 million during the fourth year of the demonstration, but much of the money came from state and federal sources. Local revenue was quite limited--one of the factors that prompted the superintendent to secure voucher funds in the early 1970's.[3] The federally funded demonstration was attractive, for it was expected simultaneously to increase district revenues and to provide a mechanism for implementing two of the superintendent's key interests: decentralization and increased parent participation in the schools.

These interests were promoted through the creation of mini-schools, instructional entities that provided for educational options within the school and competed for students and resources. The mini-schools also served as the basic organizational vehicle through which teachers participated in decentralized decisionmaking regarding curriculum, staffing, resource, allocation, and other matters.

By the fourth year of the voucher demonstration, 14 of the district's 23 schools had chosen to organize mini-schools, creating 51 nominally distinct programs. Structurally, mini-schools were organized around either a specific instructional approach or a subject emphasis, and could be classified into five different program types: Basic Skills, Individualized, Bilingual Bicultural, Learning By Doing, and Fine Arts.[4]

---

[3]The assessed valuation per student in 1972 was less than one-third of the state average, while the tax rate was double that of the median rate in California. In 1971-72, more than half of the district's funds came from state sources (one-third was the average for other districts in the state). Per pupil expenditure was $850. See Bass (1978) for further information.

[4]Chapter 2 describes mini-school types in detail. This classification is our own; it differs from that used by the district to describe the available program options.

Each voucher school contained an average of three of these different program types, which coexisted in a single school building.

In contrast, nonvoucher schools remained more conventional in their decisionmaking strategies and instructional organization. The basic organizational structure in nonvoucher schools consisted of the traditional classrooms within schools, rather than classrooms within mini-schools within schools. Without the basic vehicle for decentralized decisionmaking (the mini-school), decisionmaking authority in the nonvoucher schools tended to be more centralized than in the voucher schools (Rasmussen et al., 1977). Another potential source of divergence between voucher and nonvoucher schools was instructional orientation (e.g., subject emphasis, teaching method). While voucher schools claimed instructional orientations based on differing mini-school types, nonvoucher schools were typically described as "traditional." However, since the familiar concept of the "traditional" classroom is not well defined, a major task of the diversity study was to determine the ways in which educational practice differed between voucher and "traditional" nonvoucher classrooms.

POLICY CONSIDERATIONS

From a policy perspective, the answers to the research questions posed above bear on the issues of whether educational diversity can be implemented by an alternative system like the one tried in Alum Rock, and if so, along what dimensions. The research essentially approaches policy concerns in terms of feasibility, seeking to determine whether substantial program differentiation can be achieved by the end of a reasonably long implementation period. The research further attempts to determine whether the various educational programs generated are empirically distinguishable, and whether the distribution of classrooms among programs reflects the pattern of educational alternatives intended to result from the demonstration. Finally, by developing multidimensional program descriptions and interpretive variables, the investigation will provide researchers, practitioners, and policymakers with useful ways of characterizing classroom learning environments and processes.

PROCEDURES

Research questions and related policy issues were pursued primarily by means of multidimensional scaling of structured observation data collected in third- and fifth-grade demonstration and comparison classrooms (n = 74). Supplementary data were provided by teachers' responses to a questionnaire intended to probe areas not easily susceptible to observation. Groundwork for this research had been carried out in two previous pilot efforts. A limited range of observational data was obtained for a small sample of classrooms (n = 20) in spring 1974, in order to develop an analytic procedure for constructing and interpreting a model of classroom diversity. In spring 1975, a larger number of classrooms was observed (n = 40) in a great many respects, in order to identify a set of variables potentially capable of discriminating classrooms and to develop methods for pooling, combining, and scaling them. These pilot investigations, together with a review of relevant research literature, supported the approach to the study of classroom diversity undertaken in spring 1976.

As we have noted, the main study question was whether or not empirical groupings of classrooms could be found that were roughly coincident with mini-schools or with conceptually coherent sets of mini-schools. Secondary interest was directed at the amount of variation to be found among nonvoucher classrooms relative to that among voucher classrooms. To determine whether empirical grouping was a result of factors associated with schools per se, with mini-schools, or with individual classrooms, it was necessary to obtain a fully replicated sample of classrooms. That is, in the voucher sample, each school had to be represented by at least two mini-schools, which in turn had to be represented by at least two classrooms (see App. C). This procedure eventually generated a set of 40 voucher classrooms representing 19 different mini-schools housed in 8 schools. It constitutes the principal analytic sample for the study. The comparison sample contains 34 third- and fifth-grade classrooms representing the nine nonvoucher schools. The data, then, are hierarchically organized, the voucher sample being doubly nested (classrooms within mini-schools within schools), and the comparison sample nested only once (classrooms within schools).

Variables for characterizing sample classrooms were drawn from three sources. First, we consulted mini-school self-descriptions published by the Alum Rock school district to aid parents and students in making enrollment choices. These publications established the salient features by means of which the several educational programs perceived and distinguished themselves. Second, it seemed necessary to consult other educational literature for variables thought to represent important aspects of learning environments and thus to enable discrimination among classrooms. Major sources on which we relied included the SRI Follow-Through observation study (Stallings and Kaskowitz, 1974), White's review of evaluations of compensatory education programs (1973), Walberg's collection of evaluation studies (1974), and a recent British typology of elementary school classrooms (Bennett and Jordan, 1975). Finally, we drew heavily on our own pilot experience in Alum Rock.

Most of the variables chosen for investigation were operationalized in terms of observational measures. The resulting instrument (App. A) comprised five sections. First, a Classroom Roster to be completed with the teacher's help at the beginning of each observation period provided information about relatively stable aspects of the classroom social environment (e.g., number of students present). The Physical Environment section next generated an inventory of features of the classroom physical environment. A Five Minute Observation form recorded information about basic diversity and individualization factors (e.g., grouping, location, tasks, materials, and the like) as well as about interaction styles and other behavioral variables. Four such forms were completed consecutively to yield twenty-minute chunks of observation data, while the entire observation period involved four twenty-minute sequences representing different curricular areas (reading, mathematics, one other academic area, and one nonacademic area). The fourth section of the instrument, the Classroom Log, established a descriptive account of the classroom in the form of brief objective statements about teaching style, management style, and student behavior. Observers were instructed to attend to target events throughout the observation day, completing the log at its end. Finally, a set of

Summary Scales obtained the judgment of an experienced observer regarding the extent to which generic dimensions of self-avowed mini-school program orientation were manifest in a classroom (e.g., emphasis on multicultural themes, emphasis on basic skills, and the like).

Each classroom in the study was observed by two observers on two different occasions. A total of 14 observers required one month to complete this task. After the collection of observation data, supplementary data were obtained by means of a teacher questionnaire (App. A), which served two main purposes. First, it elicited teachers' perceptions of the learning environments they had established. Second, it sought information not readily obtained through observation (e.g., hours per week spent in team teaching, frequency of testing, and the like). Construction and interpretation of a model of classroom diversity relied on both data sources, although observational information was given primary emphasis.

ANALYSES

The goal of analyses undertaken on these datasets was to construct and interpret a spatial model of the variation among educational alternatives in Alum Rock, where proximity and distance among units would represent similarity and dissimilarity among classrooms. The result would permit determination of the existence and extent of diversity. Further, investigation of its structure would indicate whether such diversity was a function of mini-school organization. To meet this goal, a three-step analysis plan was carried out:

- First, data reduction involved pooling of observations over occasions and then pooling of variables to establish a manageable basis for subsequent analytic steps. From these, a set of "diversity" variables were selected, which were seen as reflecting necessary, if not sufficient, conditions for the practical existence of classroom diversity. The five diversity variables were all low-inference, high-reliability items from the Five Minute Observation and served as the foundation

for the spatial model. Other observation and self-report variables formed scales to be used in the interpretation effort.

- Second, sets of diversity variables were taken as constituting profiles for each classroom. From them, Euclidean distances between each pair of classrooms were computed. The resulting matrix of between-classroom derived distances provided an appropriate basis for analysis by nonmetric multidimensional scaling, yielding a spatial configuration with classrooms represented as points. The dimensionality of the space, within the inherent limitation given by the number of variables in the profiles, was determined by considerations both of goodness of fit of model to data and of interpretability. A three-dimensional solution emerged.
- Finally, an interpretation of the three-dimensional configuration of classrooms was developed by introducing into the model independent but theoretically relevant information. That is, from the set of observational and self-report variables not used in the construction of the spatial model, a subset was chosen that had some prima facie interpretive relevance. These variables were fitted as vectors in the three-dimensional space. Their relationship to the coordinates as well as to the configuration of classrooms then served as the primary guide for interpretation.

In our opinion, this approach represents a contribution to the research literature treating objective measurements of classroom organization and process, and provides a vehicle for detecting and portraying variation among learning environments over the range of variables studied.

LIMITATIONS OF THE STUDY

The investigation was subject to numerous constraints that limit its external and internal validity. With respect to external validity, it seems clear that since the variables measured are reasonably descriptive of classroom organization and process in general, limitations on

generalizability come not from the variables employed, but from the sample itself. The study setting is, in the strictest sense, unique. All classrooms were drawn from a single school district participating in a federally sponsored demonstration, which is why this investigation must be regarded as a case study of diversity. Given the basic nature of the variables themselves, what is most in doubt is the generalizability to other settings of the general findings regarding extent of diversity. These findings indicate that (with exceptions to be discussed in detail later), voucher demonstration classrooms did not manifest noteworthy differentiation along the dimensions explored. Nonvoucher classrooms showed essentially no variability beyond that exhibited by the mass of voucher classrooms. Consequently, only to the extent that nonvoucher classrooms selected for comparison purposes are taken as representative of traditional elementary education would we expect our results to be replicable elsewhere.

With respect to internal validity of findings, the most serious limitation of the study concerns its use of derived rather than directly judged distances (or similarities) for construction of the model. Use of distances computed from preselected variables necessarily restricts the assessment of classroom similarity to dimensions specified a priori. These dimensions, however, may not coincide with or exhaust the relevant ways in which learning environments could be assessed as similar or dissimilar. Thus, the resultant configuration may not reproduce the sort of arrangement that would be produced by judgments from experts familiar with the classrooms. Moreover, without overall direct similarity estimates, it is not possible to perform an independent check on the adequacy of the derived distance configuration based on profiles constructed from sets of observations. Limitations of time and funding precluded the possibility of convergent measurement of similarity by direct estimation. A second internal constraint has to do with lack of adequate data from the repeated observations. For reasons discussed further below, data from the second round of observations were incomplete and were not included in the final analysis. Therefore, the true extent of variation on the measured dimensions over time is not known. However, exploration of second-round observation

data suggests that it would not be sufficient to alter in any meaningful way the findings from analyses of the first round of observations.

A final potential constraint--but one that evidently did not materialize--has to do with attitudes of teachers toward the study. Participation in the study was mandatory for all teachers whose classrooms were selected for the sample, and many were not pleased at the idea of being observed. Moreover, because teacher layoffs were anticipated at the end of the school year (a matter discussed at greater length below), they were probably apprehensive over being evaluated. Under such circumstances, it would not have been surprising to find that teachers altered their usual behavior in unknown ways because of the presence of the observer. For a number of reasons, we do not believe this happened. First, after the study was mandated, steps were taken to make the procedures as nonthreatening as possible. Second, teacher cancellations on scheduled observation days were few, and these sessions were rescheduled without difficulty. Last, observers--who were themselves substitute teachers from a neighboring district--reported that teachers of observed classrooms did not seem either hostile or apprehensive and, after the first few minutes, hardly seemed to notice their presence. We therefore believe it likely that teachers' attitudes toward being observed had little effect on the outcomes of the diversity study.

Chapter 2 below treats the conceptual framework for the investigation, describing in detail the nature and origin of the variables on which classroom profiles were based. Chapter 3 presents the two instruments in terms of which the variables were operationalized, and gives a complete account of sampling and data collection procedures. The construction of a spatial model of classroom diversity in Alum Rock is the focus of Chapter 4. Building on Chapter 4, Chapter 5 provides an interpretation and Chapter 6 a discussion of the results.

Chapter 2

CONCEPTUAL BACKGROUND

During the fourth year of the demonstration there were more than 50 mini-schools in Alum Rock. The availability of over 50 programs in which to enroll makes a *prima facie* case for the existence of diversity in schooling, since it would be difficult to argue that participating classrooms did not differ in any respects. However, even a multitude of differences among classrooms are no guarantee of significant diversity. Rather, the issue of significant diversity turns on whether classrooms differ substantially along dimensions thought to be important constituents of the learning environment, and whether such differences can be systematically integrated to provide a set of coherent alternatives.

APPROACH TO THE STUDY OF CLASSROOM DIVERSITY

In inquiring whether diversity existed in Alum Rock classrooms, and if so, what it consisted of, the present study was driven more by the intent to see what was there than by any specific expectation about what types of educational environments would be found. As we have noted, the investigative approach was not one in which the investigator attempts to determine whether a set of *a priori* program models has been implemented. Such an effort would be broadly akin to pattern matching, and is exemplified by Stallings and Kaskowitz's (1974) approach to the Head Start--Planned Variations project. A major goal of that study was to assess the extent to which varied compensatory education program models had been implemented in different sites. For that purpose, critical features of each model were specified in advance and classroom observations were employed to determine their presence or absence at intended implementation sites. Observing a site and classifying it under a given model, then, is primarily an exercise in empirical verification.

In contrast, the study reported here is primarily an attempt at typology induction, since it was not known in advance what features,

if any, distinguished educational programs. It was assumed, nevertheless, that the establishment of diversity depended on finding some basic dimensions of program differentiation. Consequently, the difference between this approach and the former one lies in the source of the dimensions that are said to define classroom types. In the present case, locating and describing those dimensions constitutes an important projected result of the study rather than an antecedent condition. The research began with a large set of variables theoretically and empirically relevant to the determination of classroom diversity, from which underlying dimensions of differentiation could be extracted.

Given the nature of the study, then, the selection of initial variables assumed considerable importance. The research sought to discover which characteristics of classrooms among all those studied generated meaningfully different educational choices. The legitimacy of the responses made to this question depends on having selected in advance a sufficiently comprehensive and relevant set of variables, and having measured them adequately. Should the investigation not satisfy these conditions, it could always be claimed that obtained outcomes (whatever they might be) would either differ more sharply or would correspond more closely to the real situation had more appropriate variables been included or had different assessment methods been employed.

For these reasons, the discussion below gives careful attention to the way in which variables were selected for the study. That selection was guided by the following considerations:

- First, mini-school self-reports were regarded as a face-valid source of potential program dimensions. That is, the mini-school self-descriptions revealed how programs viewed themselves, the terms in which they characterized their educational offerings, and what they took to be their salient differences. No attributes that played a part in these descriptions could be excluded from the study.
- It also seemed necessary to consult other educational literature for variables thought to represent important aspects of the learning environment that would enable discrimination among

classrooms. To look at educational alternatives exclusively in terms of how mini-school practitioners construed them would risk neglecting classroom differences that arose as a function but not as an explicit aim of mini-school program development. For this purpose, we especially sought research literature related either to establishing program diversity or to evaluating diverse programs.

- Next, variables were sought that could be assessed primarily by observational methods, for two reasons. First, according to Goodlad et al. (1970), "many of the changes we have believed to be taking place in schooling have not been getting into the classrooms." The Goodlad study of educational innovations indicated that while teachers and principals believe novel programs have been implemented and perceive their own behavior accordingly, classroom observations find few departures from traditional practice. Consequently, to judge the extent to which distinct educational alternatives have in fact been implemented, observation seems required. Second, Berman and McLaughlin (1977) found that even in cases of successful innovation, a program is rarely implemented in precisely the manner originally specified (cf. Mason, 1973). Thus, to establish the nature of educational options as they evolved during the course of implementation, observation seemed an important vehicle.

From an initial abundance of classroom characteristics generated in these ways, a subset was chosen for which: (a) applicability and relevance to the Alum Rock classroom setting were clear; (b) observational reliability was established; and c) interclassroom variability was obtained. Such determinations were made on the basis of extensive pilot exploration. These considerations, along with the classes of variables they generated, are discussed in order below.

Mini-School Self-Reports

Each spring the Alum Rock school district published a directory of mini-schools for the forthcoming academic year, so that parents and

students participating in the demonstration would know what choices were available to them. The Voucher Choices directory for the academic year during which the diversity study was conducted provides accounts of 50 programs (Sequoia Institute and Voucher Demonstration Project, 1975). These descriptions constituted summaries of information supplied by mini-schools in response to an interview concerned with classroom organization (e.g., amount and kind of teacher direction of activities, degree of student choice), instructional materials used in reading and mathematics, special events (e.g., multicultural activities, out-of-class activities, arts and crafts), and parent involvement.

In addition, to facilitate comparison of programs, mini-schools were grouped under "types" on the basis of presumed similarity of orientation. The "types" supplied by the Voucher Demonstration Project cross two quite different classification principles: subject emphasis (e.g., Basic Skills) and learning process (e.g., Learn by Doing). Consequently, they cannot form the basis of an a priori program typology for confirmatory research purposes. This is not to say that program self-descriptions are without value. Rather, their value is more likely to lie in suggesting variables for consideration than in supplying an initial categorization. The five mini-school "types" elicited from the program directory are briefly described here to indicate the kinds of differentiating dimensions they propose. (Listings that excluded the grade range sampled in this study have been omitted from the discussion.)

Basic Skills entries are the most numerous in the directory. These mini-schools strongly emphasized reading, language arts, and mathematics, and tended to describe themselves as teacher-centered. That is, learning was said typically to involve the entire class, with the teacher explaining, demonstrating, and conducting drills. However, work was individualized for students who were either behind or ahead of the rest of the class. Activities appeared to be highly structured, with student-student interaction at a minimum except during nonacademic periods. Use of learning centers and traditional materials was reported. Student-teacher ratio on average was 16:1.

Individualized mini-schools, the second most numerous listing, were also basic-skills-oriented, making use of traditional instructional

materials and learning labs. However, their main concern was to provide a classroom environment in which students worked by themselves at their own level and pace, using the particular materials that best suited their needs. Consequently, there were said to be no class lectures or drills, and no occasions for students to work competitively. A mean student-teacher ratio of 17:1 was claimed.

Third in frequency of occurrence were Bilingual/Bicultural mini-schools. These programs focused on language arts in a culturally pluralistic context. Basic skills instruction was said to proceed in the students' first language (either Spanish or English), with considerable emphasis on learning a second language (English for Spanish-speaking students, and Spanish for English-speaking students). Some nontraditional instructional materials were employed. An attempt was made to individualize instruction as well as to encourage cross-cultural peer interaction. An average 15:1 student-teacher ratio was reported.

Next in order of frequency were Learn By Doing mini-schools, whose aim was self-teaching of basic skills through direct experiences in an active, "doing" environment. Learning was reportedly mediated by games and puzzles as well as by practical activities such as cooking, camping, ecology projects, and the like. Loosely structured small-group concurrent tasks were apparently preferred, with classroom processes tending to be student-centered and teachers adopting a facilitator role. These mini-schools averaged a 14:1 student-teacher ratio.

Finally, a small number of mini-schools attempted to mediate academic learning through Fine Arts (dance, music, drama, graphic arts). Emphasis was on creative activity in individual or small group projects, with nontraditional learning media available. The modal student-teacher ratio reported was 15:1.

Among variables emerging in the self-descriptions summarized here, some are relatively palpable and structural in nature, and seemingly susceptible to unambiguous interpretation. Class size, student-teacher ratio, composition of working groups (whole class, subgroup, individual), type of instructional material, and the like are included in this set. In contrast, classroom management and process accounts tend to be more vague. As the directory remarks, the mini-schools vary from

"traditional" to "open," often involving elements of both. However, such terms are especially in need of operational refinement. As Bennett and Jordan (1975) point out in the introduction to their typological study, not only are these terms "often ill-defined," but "The same terms are used to refer to different aspects of the classroom situation, and different terms to refer to essentially the same aspects." Because mini-school self-reports fall heir to these difficulties--in fact, they do not claim to describe classroom similarities and differences precisely and exhaustively--relevant research literature was consulted in order to clarify and supplement the descriptors they supplied.

Background Literature

Our investigation of research literature for variables capable of discriminating learning environments was guided by two principles. First, we sought studies aimed either at establishing an educational program typology or at confirming that a program model(s) had been implemented, rather than outcome studies. This constraint significantly limited the extent of the survey. Second, we gave highest priority to research involving low inference dimensions of educational practice, since these would be most readily given operational form. The review generated a great number of research efforts focused at one or a small number of interrelated classroom variables such as individualization (Cohen, 1973), participation (Risley and Cataldo, 1973), cooperation and competition (Johnson et al., 1976), self- and other-initiated learning (Condry, 1977), teacher approval and disapproval (Heller and White, 1975), and the like. However, of greater use for our purposes were more broadly conceived studies intended to provide systematic multivariate accounts of the learning environment as a whole. Extremely valuable in this regard were the Follow Through implementation study conducted by Stallings and Kaskowitz (1974), and the typology induction study of Bennett and Jordan (1975). Literature reviews giving explicit attention to the problem of determining and organizing basic dimensions of learning environments were also very helpful (White, 1973; Brandt, 1973; Siegel and Siegel, 1967; Walberg, 1974). For convenience, the results of the investigation of background literature are grouped

under three topics: classroom setting; classroom processes; and teacher student roles.

Classroom Setting

With respect to classroom setting (a term borrowed from Siegel and Siegel, 1967), virtually all sources consulted treated the number and distribution of persons. That is, class size and student-adult ratio were taken by most sources as basic descriptive variables, along with number of classroom aides, resource persons, and volunteers (cf. Stallings and Kaskowitz, 1974; White, 1973; Anderson and Walberg, 1974). Equally important was the question of grouping, i.e., whether the class works as a whole, whether it is divided into large or small groups, and whether students work individually or with a tutor. Grouping patterns were a dimension of primary interest for Stallings and Kaskowitz (1973), Cohen (1973), and Brandt (1973) typological study.

In addition to an account of the distribution of persons in the classroom, information about the physical environment is often included with setting variables. Such information may concern "the physical characteristics of the space" (Siegel and Siegel, 1967), that is, what kinds of work locations are available (seating at desks, learning centers, activity centers, and the like) and whether the student or teacher makes the decision where to work (cf. Bennett and Jordan, 1975; Stallings and Kaskowitz, 1974). Closely related are inventories of work materials, typically included in descriptions of classrooms. Cohen (1973) contends that a diversity of materials is critical for individualization of instruction, a view supported by many other sources (cf. review in White, 1973). On the other hand, as Brandt (1973) points out, "there is no widely accepted taxonomy of classroom materials." It was concluded that, while a report of task materials would be desirable, the items included for study should be based on pilot experience in Alum Rock together with mini-school self-descriptions, and that the most significant index might well be not the materials themselves, but whether or not a multiplicity of materials were in use relative to a given task.

A final group of setting variables concerns instructional emphasis. Instructional emphasis includes, first, curriculum, the kinds of

subject matters pursued by the students throughout the class day. Here most sources mention not only the content of activities, but also whether the student or the teacher chooses the task to be done at a particular time and whether diverse areas of learning are integrated or taught separately (Bennett and Jordan, 1975; Brandt, 1973; Stallings and Kaskowitz, 1974; White, 1973). Closely connected to instructional content is instructional strategy, a construct that deals with the mode of learning. During any task period, a subject matter may be approached by question/answer drill, discussion, experimentation, board work, teaching machines, and so on (Brophy, 1975; White, 1973; Stallings and Kaskowitz, 1974). Although this review produced a potentially large number of setting variables, White (1973) suggests that virtually all of them can be viewed in terms of amplification or reorganization of traditional classroom services.

Classroom Process

Next to be discussed are classroom process variables. While the setting variables treated above primarily involved descriptions of the physical and social environment, process variables attempt to portray how activities are conducted within that environment. In general, the literature affords greater consensus with respect to determining basic setting variables, perhaps because process variables tend to be more inferential. Nevertheless, there is some overlap among these variable groups. For instance, while type of task and work location were included in the discussion of setting variables, another important feature of the description was who chose what to do and where to do it--the teacher or the student. For the purposes of this review, the question of who assumes responsibility for leading or directing a given activity will be treated as a classroom process dimension. Most of the sources consulted mentioned one or more kinds of classroom choices. In addition to those already noted, selection of task group, long-term seat assignment, and choice of the learning material and mode were also cited (cf. Brand, 1973; Bennett and Jordan, 1975; Condry, 1977). Special attention is given by Cohen (1973) to the pacing of work, since, she contends, selfpacing is critical for diversifying and

individualizing activities. Finally, individualization of work receives particular emphasis by Bredo (1975).

Besides work leadership, another frequently encountered set of process variables involves the nature and extent of interaction during learning activities. Virtually all sources treated both teacher-student interactions and student-student interactions. With respect to teacher-student interactions, salient dimensions were how often such interactions occur; who is the initiator; whether the teacher interacts with individuals or groups of students; and, when students are divided into groups, whether the teacher remains with the groups or leaves them to work independently (cf. Stallings, 1973; Brophy, 1974, 1974; Brandt, 1973; Siegel and Siegel, 1967). In relation to peer interactions, a basic question was whether or not they are permitted during learning activities. Given that there are tasks during which student interaction permissibly occurs, further distinctions are made concerning whether the task is designed to require group interaction, and whether interactions are cooperative, competitive, or neutral in their goal structures (e.g., Johnson et al., 1976; Stallings, 1974; Cohen, 1973; White, 1973).

While the sets of process variables just discussed account for most of the consensus in the reviewed literature, two additional dimensions of classroom process are worth noting here. First, several sources cited student attention to or involvement in tasks as an important aspect of the learning environment. In particular, this dimension has been the focus of investigation for Risley and Cataldo (1973), who contend that extent of participation in appropriate activities is a critical index of the quality of classroom settings. Further, Brophy's (1975) factor analysis of five well-known classroom observation systems yielded a first factor interpretable for student task participation, providing empirical support for the significance of this construct in characterizing learning environments. The remaining classroom process domain of interest concerns performance feedback. A great deal of attention is given to whether or not students receive feedback from their teacher during learning activities. Assuming that they do, it is relevant to inquire how often feedback is given, of what sort it is

(comments, grades, tokens, and so on), and whether it is positive or negative (Bennett and Jordan, 1975; Heller and White, 1975; Condry, 1977; Glass, 1974; Stallings and Kaskowitz, 1974; Brandt, 1973).

Teacher and Student Roles.

The last group of variables considered here involve teacher and student roles, a phrase adopted to stand for a loose and open-ended collection of behaviors, styles, and attitudes characteristic of teacher or student activity in the classroom environment. While most of the sources reviewed mentioned some variables appropriately included under this heading, there was no striking consistency of items across sources nor consensus regarding their importance or desirability. It should be noted that the boundary between classroom process dimensions and descriptions of student-teacher roles is rather arbitrary, and many of the latter could equally well have been grouped with the former. In general, however, it seems that the role variables either refer to very specific classes of events (in contrast with process dimensions, which are typically applicable across a wide range of activities), or else involve higher-inference constructs. The discussion of role variables treats first, characterizations relevant to teachers, then characterizations related to students, and finally characterizations that involve both.

Among teacher role variables, the most frequently mentioned is the teacher's questioning style. Whether open- or closed-ended responses are sought by the teacher was a defining item for the third factor in Brophy's (1975) factor analysis of classroom observation scales. In addition, this variable was helpful in differentiating program types for Stallings and Kaskowitz (1974) and for White (1973). Other aspects of questioning style seen as important were whether questions are aimed at the group or at individuals (Stallings and Kaskowitz, 1974), and whether or not recitations are elicited in patterned order from students (Bennett and Jordan, 1975). A seemingly related aspect of teacher behavior concerns whether instructional methods tend to be direct or not. The term "direct" is used by Brophy (1974) to describe instructional situations in which the teacher holds the floor and presents information

to students by means of lecture, demonstration, explanation, and the like. Direct teaching is regarded by Brophy (1974) as an important role variable, and it also appears (although not by that name) in both Brandt (1973) and Stallings and Kaskowitz (1974). A final teacher behavior dimension given substantial attention is the teacher's use of student ideas, a variable treated by Flanders (1970), Brophy (1974), White (1973), and Rosenshine and Furst (1971). In addition, Walberg (1974) and Rosenshine and Furst (1971) include the extent to which the teacher related subject matters to student ability level and interest, while White (1973) mentions the teacher's use of the student's own language and culture as a medium of instruction.

Another class of teacher role variables pinpoints expectations, attitudes, or traits. Often cited in this category is the teacher's work orientation, described variously as: The teacher has rules and will enforce them; the teacher expects a great deal of work to be done in class; the teacher is well organized; the teacher is task oriented (cf. Brophy, 1974; Walberg, 1974; Rosenshine and Furst, 1971). Other teacher dimensions mentioned include warmth and enthusiasm (which define the second factor in Brophy's 1975 factor analytic study of classroom observation scales), clarity, cognitive level, and quality of presentations (cf. Brophy, 1974; Siegel and Siegel, 1967; Rosenshine and Furst, 1971; Bargen and Walberg, 1974).

Student role dimensions are less frequently the subject of classroom-level studies (especially when the research aims at eliciting or modifying responses in instructional settings). Paralleling the interest in teacher questioning is a concern for students' question-asking style. Several sources cited as important the frequency with which students initiate questions, as well as whether the questions tend to be objective or open-ended (Stallings and Kaskowitz, 1974; Brand, 1974). A more generic concern was whether students independently seek help when they need it (Brophy, 1974), and whether they engage in self-teaching and self-evaluation (Condry, 1974; White, 1973). In addition, several sources cited as an important student role dimension the extent to which students participate in discussions or extended exchanges (Stallings and Kaskowitz, 1974; Flanders, 1970; Prescott et al., 1967).

The literature surveyed generally yielded few student characteristics that would form a counterpart to the teacher role characteristics discussed above. Siegel and Siegel (1967), however, do include student creativity, educational set, and specific task motivation as potentially important descriptors for a multivariate picture of the classroom setting. Further, Anderson and Walberg (1974) cite perceived cohesiveness, friction, clique structure, satisfaction, disorganization, and apathy as significant factors in students' experience of the classroom. However, it is not clear how these experiences of classroom climate map onto the objective properties of that environment.

A final small set of descriptors have implications for both student and teacher behaviors, being aimed at kinds of classroom activities. The use of games or "fun" activities for learning is cited by Brophy (1974), while White (1973) treats the occurrence of discussions about feelings, attitudes, and behaviors toward self and others as a discriminator of types of classrooms. Both Brophy (1974) and Risley and Cataldo (1973) give some importance to the smoothness of transition from activity to activity versus time spent waiting. Last, Bennet and Jordan's (1975) classroom research yields assessment method as an independent descriptive factor. Included here are such items as tests, in-class assignments and homework to be completed, whether or not these are corrected, and if so, by whom. Undoubtedly, many more dimensions of learning environments could be added to this account, which makes no claim to completeness. The discussion here surveyed only the educational literature most relevant to our research purposes, selecting from it classroom descriptors representative of greatest consensus among sources. Their potential applicability for differentiating Alum Rock classrooms and their susceptibility to reliable operationalization primarily by observational measures were also considered. There considerations are treated in more detail below.

REFINEMENT AND OPERATIONALIZATION OF VARIABLES

As we have indicated, final selection of variables for the diversity study turned on issues of appropriateness to the measurement context. The varied and numerous classroom descriptors generated by

the literature review together with the mini-school self-reports were examined from two perspectives: which of them could be reliably operationalized, preferably by observational methods; and which of them were useful for discriminating Alum Rock classrooms.

The rationale for employing observational measures whenever possible has already been discussed. Briefly, it is likely that: (a) programs differ in ways not apparent to participants; (b) nominally distinct programs may be similar and nominally similar programs, distinct; and (c) previous research suggests that participants do not make accurate judgments about the extent to which a program has been implemented. A secondary objective for using observational measures was the contribution the study might make to understanding elementary education settings ethologically. A naturalistic study of a sizeable number of classrooms would provide a descriptive data base potentially useful for future research.

Having established the desirability of observational assessment, we then sought guidelines for the operationalization of variables of the sort discussed in the preceding parts of this chapter. In general, guidelines for the observational study of classrooms are easy to find. A great number and variety of observation systems have been developed in recent years. Among these, a substantial subset develop procedures that are focused at the individual rather than the classroom or group level and that code individual behavioral events as they occur (such as single sentences or single acts). Such systems were consulted for whatever light they might shed on the general problem of designing an observation procedure. However, they provided little in the way of specific models for the present study (e.g., Foa, 1961; Flanders, 1970; Waxler and Mishler, 1967; Fine and Zimet, 1956; Longabaugh, 1963; Moustakas et al., 1969; Brophy et al., 1975; Risley and Cataldo, 1973; Stallings and Kaskowitz, 1974; Prescott et al., 1967; Calfee and Hoover, 1974; Rosenshine and Furst, 1973). Within the latter group, we relied most heavily on Risley and Cataldo (1973) and Stallings and Kaskowitz (1974), along with Calfee and Hoover (1974) and Prescott et al. (1967), for specific guides to operationalizing variables of interest.

Final selection of variables for the diversity study, however, was most heavily influenced by pilot work undertaken by project staff in spring 1975. In that effort data were obtained from 40 volunteer classrooms, representing both voucher and nonvoucher schools, during two twenty-minute reading periods. The observation instrument contained over 150 variables, conceived as reflecting some 18 underlying or more generic dimensions of classrooms that had been frequently referenced either in the mini-school self-reports or the literature review. The data were treated in several ways. In some instances, transforms were examined as well as original scores (for example, in addition to a count of materials used in a task, mean number of task materials and standard deviation or coefficient of variation about the mean were analyzed). Second, intercorrelations among items were investigated in order to determine which variables could appropriately be summed. Third, distributions of classrooms along original variables, transformed variables, and summary variables were explored to see which dimensions showed most promise for discriminating classrooms. Finally, a subset of variables potentially most useful for the diversity study were identified and subjected to an exploratory factor analytic procedure (ACOVS, Joreskog, Gruvaeus, and VanThillo, 1970) intended to suggest higher order dimensions that might ultimately define a multidimensional classroom space. The results of this effort are described below, organized in terms of the 18 hypothetic factors grouped into four categories that conceptually unified the original set of pilot observation variables.

Physical Setting

Physical setting accounted for six of the hypothetic factors. It should be recalled that mini-school self-descriptions typically reported features of the physical environment as important aspects of their program, and that the surveyed literature frequently mentioned environmental inventories (included under classroom setting variables in the review above). However, the lack of a widely accepted environmental taxonomy for classrooms suggested that we try a number of items in the attempt to develop an appropriate inventory for the present project. Accordingly, we investigated groups of items representing: (1) student-usable

equipment (e.g., tape casettes and earphones, self-teaching machines); (2) teacher usable equipment (e.g., overhead projectors); (3) environmental comfort (e.g., carpeting, lighting); (4) diverse nonacademic materials (e.g., maps, puzzles, posters, student art displayed, cultural displays); (5) open or closed physical arrangements (e.g., desks in rows or clusters, rooms divided into separate areas, openings between classrooms); (6) traditional academic materials (e.g., sets of texts, written work displayed, science exhibits, rules posted). Over 40 physical setting variables were examined.

The variables representing the first three hypothetic factors were in many instances susceptible to ceiling or floor effects. Those variables for which a reasonable distribution was obtained intercorrelated at a low positive level. Summing these to form a single "equipment" index seemed the best course for discriminating sampled classrooms. Among items from the fourth factor, only the presence of posters and cultural displays exhibited interpretable patterns of correlation and exhibited reasonable between-classroom variance. These items were strongly intercorrelated and, for future instrumentation purposes, would be treated as a single environmental dimension. With respect to physical arrangements, only the use of clustered desks and room dividers seemed able to differentiate classrooms. While such arrangements were positively correlated, use of room dividers was also associated with the summed equipment index and with presence of cultural displays. Consequently, these were retained as separate items. Finally, the host of items representing a traditional academic environment did not comprise an interpretable whole. While some variables were positively associated as expected (e.g., sets of texts and the presence of books other than texts), others yielded an unexpected negative intercorrelation (e.g., displayed written work, sets of texts, and science exhibits). It was decided that the variables were either poorly defined, inappropriately scored, or else not a relevant set for the Alum Rock classrooms.

Social Setting

Social setting variables comprised items reflecting both social environment and classroom processes (see setting and process variables

in the review above), and were conceived in terms of three hypothetic factors. (1) Differentiation was represented by the number of groups into which the class was divided, mean group size, coefficient of variation about mean group size, and a logarithmic representation of the distribution of group sizes. In addition, differentiation included indices of the different kinds of materials, work areas, and activities with which the groups were involved. (2) A set of variables reflecting routinization included the extent to which students choose what to do, change or stability in the composition of groups, and students' freedom to join or leave groups. Also examined as part of this set was the extent to which students are self- or other-paced (means, standard deviations, and coefficients of variation were employed). (3) A final construct, individuation, was reflected by such variables as the number of instructors, the adult-student ratio, the use of adult and student tutors, the number of adults working with each group, the amount of time spent teaching by aides, and the amount of adult movement.

Among the social setting and process variables, those regarded as representative of differentiation seemed strongest. That is, they formed a cohesive set capable of discriminating classrooms. The logarithmic representation of sizes of classrooms was especially sensitive to between classroom differences, providing a better index than either mean group size or coefficient of variation about that mean. However, relationships between these and other variables were not always straightforward. For instance, number of adults was positively associated with use of math and reading centers, but all three were negatively correlated with the use of peer tutors and with many of the differentiation variables (such as diversity of tasks and materials). In addition, the number of adults--especially when the adults move about a great deal in the classroom--was negatively associated with student choice and student movement. It seemed clear that individuation and differentiation should be regarded as independent constructs. Moreover, it seemed possible that having relatively large numbers of active adults in the classroom might preempt opportunities for diverse student-initiated activities.

Routinization variables, in contrast, formed a less solid potential factor. Several routinization items were strongly associated with

either differentiation or individuation, leading us to question whether it should be regarded as a separate construct. Among the routinization variables, pacing was most independent and most capable of discriminating classrooms (specifically, standard deviation of mean extent of other-pacing seemed the best measure). Observer comments suggested that a better operational definition of the pacing variable might produce better results.

Teacher Role

Teacher role variables (see the discussion of student and teacher roles in the review of literature) provided the content for a number of tentatively identified classroom factors. While mini-school self-reports had not emphasized such descriptors, they comprised a substantial, although not highly consensual, proportion of the reviewed classroom literature. Thus, it seemed appropriate to pilot a variety of teacher role variables in order to arrive at a viable set for the final instrument. Six potential factors were explored. (1) Direct-indirect teaching played a major part in the classroom literature and was represented in the pilot study in terms of eight observer-rated items (e.g., extent of explaining, lecturing, facilitating, discussing). (2) Academic feedback was rated in terms of four items (e.g., amount positive, amount negative, amount verbal, amount nonverbal). (3) Similarly, disciplinary feedback, or feedback about personal-social behavior, was rated using the same four items. (4) Teacher effectiveness was judged by observers by means of variables describing clarity of presentations, task orientation, smoothness of transitions, degree of organization, and so on. (5) Teacher bias was reflected in items directed toward race- and sex-differentiated behavior (e.g., "Students are told how boys should be and how girls should be" or "The teacher pays less positive attention to minority children than to others."). (6) Finally, teacher management and structure data were gathered from interviews with the teachers. Questions primarily dealt with methods employed for classroom control (decision-making and evaluation procedures), and the need for and implementation of order (e.g., preassigned seats, ability-based groups).

The exploration of teacher role variable eliminated some dimensions and generated other tentative dimensions that would need further refinement before use in the diversity study. The teacher effectiveness and teacher bias measures in the main exhibited no stable properties, possibly because most of these were high inference items. However, teacher warmth and teacher energy (proposed effectiveness items) were assessed consistently enough by observers to be worth further consideration. Items intended to reflect direct-indirect teaching were also largely unsatisfactory, exhibiting virtually no variance. Exceptions were measures of teacher facilitation and teacher explanation, which showed limited potential for discriminating classrooms and which exhibited a low but significant positive correlation. The two teacher feedback dimensions held some promise. Academic feedback items were not intercorrelated. However, rated positive feedback and a summary index of total academic feedback showed reasonable between-classroom variance. Disciplinary feedback proved to be a stronger dimension, these items being intercorrelated. Again, rated positive feedback and a summary variable representing total disciplinary feedback seemed the most sensitive indicators. Most promising of all were teacher management and teacher structure items, summed to make two scales because items exhibited reasonable patterns of association.

Student Role

Student-related variables (see treatment of student and teacher roles in the background literature survey) accounted for the three hypothesized factors. (1) Student-teacher interaction was recorded in terms of its tone (positive, neutral, negative, or mixed), as was (2) student-student interaction. The latter dimension also included student cooperation items. (3) Finally, student involvement was reflected by such items as participation in ongoing activities, attentiveness, idleness, and disruptiveness. All three dimensions seemed potentially valuable for discriminating classrooms, although items appeared to need better measures. In particular, it seemed that recording interaction and participation at 10- to 20-minute intervals did not provide a sufficiently fine-grained picture of student behavior. Further, student-teacher

interaction was more closely related to teacher behavior than to student behavior, showing positive associations with rated teacher movement and energy as well as with the presence of adults in student groups.

The pilot investigation of learning environments established conceptual domains in which further operational refinement was required before viable observation measures could be taken. In addition, it yielded the 40 potentially useful variables for the diversity study listed in Table 2.1; these variables were distributed among the conceptual categories described above as follows:

- Variables 1-3 reflect classroom physical setting, the first summing kinds of equipment available for use by students and teachers and the remaining two indicating room arrangement features.
- The next group of variables reflect classroom social setting; specifically, items 4-9 were intended to represent differentiation, items 10-13 were intended to represent routinization, and items 14-20, individualization.
- Ten variables are taken from the teacher role category: variables 21-22 index direct teaching, 23-26 designate teacher feedback (academic and disciplinary). 27-28 are drawn from teacher effectiveness measures, and 29-30, from teacher structure and management measures.
- The last ten variables come from the student role category: 31-34 tap student-teacher interaction and 35-37 tap student-student interaction; 38-40 represent student involvement.

These variables were treated by means of an exploratory factor analysis for the purpose of finding more generic dimensions to be taken as predictive of the ways in which a multidimensional classroom space might be determined. The results of that analysis are also provided in Table 2.1. Each variable appears in the left column, and its factor loading (>.30) is represented by either a plus or minus sign under one of the four factors. When an item has a loading higher than .30 on two or more factors, all such loadings are indicated.

Table 2.1

PILOT FACTOR ANALYSIS

	Variable	Factor I	II	III	IV
1.	Equipment				+
2.	Clustered desks			-	+
3.	Room dividers			+	
4.	Group size (logarithmic measure)	+			
5.	Different work locations in use	+			
6.	Different ongoing activities	+			
7.	Different materials in use	+			
8.	Average number of materials per group	+			
9.	Rated materials diversity[a]	+			
10.	Extent of other pacing			+	
11.	Ability grouping[b]				-
12.	Group stability[b]			+	
13.	Students leave or join groups[a]	-	-		
14.	Adult-student ratio	+			-
15.	Number of adults	+			-
16.	Number of tutors	+	+		
17.	Rated amount of tutoring[a]	+	+		
18.	Number of groups the teacher assists			+	
19.	Number of groups other adults assist	+			
20.	Number of hours aides teach[b]	+			
21.	Teacher facilitates[a]	+			
22.	Teacher explains[a]	+			
23.	Total academic feedback		-		
24.	Positive academic feedback		-		
25.	Total disciplinary feedback		-		
26.	Positive disciplinary feedback		-		
27.	Teacher warmth[a]		-		
28.	Teacher energy[a]		-		
29.	Teacher structure	-			-
30.	Teacher management				-
31.	Student-teacher interaction (mean)	-	-		
32.	Student-teacher interaction (s.d.)			+	
33.	Student-teacher interaction tone (mean)[a]		-		
34.	Student-teacher interaction tone (s.d.)[a]			+	
35.	Student-student interaction (mean)	+			+
36.	Student-student interaction (s.d.)			+	
37.	Student cooperation	+	-		+
38.	Student participation (mean)	-	-		-
39.	Student participation (s.d.)			+	
40.	Rated student attention		-		-

[a]Does not appear in the final set of instruments.

[b]In the final set of instruments, is represented in the teacher questionnaire rather than in the observation sheet.

The analysis whose results are presented in Table 2.1 yielded four factors tentatively interpreted as reflecting diversity (factor I), classroom interaction (factor II), behavioral variation (factor III), and nontraditional environment (factor IV). Among these preliminary factors, the first two appeared to be very strong and the latter two relatively weak. It was concluded that a usable set of variables had been adequately developed to represent diversity and interaction, although recording such information at more frequent intervals for a longer observation period might strengthen them. In addition, operational definitions might be improved for certain of these items as indicated in the preceding discussion. In greater need of improvement were items representing the quality, structure, and focus of student behavior (factor III), and items representing alternative nontraditional environments (factor IV). The second set of changes was made by narrowly rereviewing the literature for suggestions and modifications in the two targeted areas. The discussion of data collection instruments that follows reflects these revisions.

Chapter 3

PROCEDURES

This discussion of procedures is divided into three major sections: a detailed account of data collection instruments, adhering closely to the conclusions drawn in Chap. 2 (copies of final-form instruments are provided in App. A); a summary of the sampling plan and its rationale (a more technical treatment of the sample appears as App. C); and data collection activities.

INSTRUMENTS

The Classroom Observation Sheet and a supplementary Teacher Questionnaire (reproduced as App. A), developed on the basis of mini-school self-reports, related research literature, and pilot investigation as described above, provided the foundation for assessing the extent to which a diversity of educational alternatives had been established in the Alum Rock demonstration. Each instrument is discussed below, with the Classroom Observation Sheet receiving greater emphasis because of its more central role in the research plan.

Classroom Roster

The Classroom Observation Sheet is composed of five parts, each representing a different kind of information. The first part, called the Classroom Roster (see App. A), describes relatively stable aspects of the classroom social setting, including some social setting variables from the pilot factor analysis (Table 2.1, items 11, 14-16). It is completed only once per occasion of observation, with the help of the classroom teacher. Modeled after the beginning section of Stallings and Kaskowitz's (1974) observation instrument (Classroom Summary Information), the roster first obtains the number of students enrolled in the class and the number present on the observation day. In addition, it inquires how many students are sent out of the class for certain instructional activities (e.g., to resource rooms, to teaching specialists). Like the Stallings and Kaskowitz summary,

the roster asks for the number of teachers, regular aides, and volunteers. To this list it adds personnel categories specified frequently in the Alum Rock mini-school self-reports (e.g., specialists, bilingual adults, older students). The roster asks for grade-level information, as does the summary, allowing for the possibility that classrooms are ungraded or have multiple grades (as reported in several mini-school descriptions). Further, the roster establishes whether students are preassigned to groups and to seats. Finally, the roster collects information about the duration of the class day and, by the end of the observation occasion, contains a record of the distribution of class time in activities. The roster, then, supplies basic information about the population of the classroom that aids in interpreting data obtained subsequently.

Physical Environment

The second section of the observation instrument represents the classroom Physical Environment (see App. A). Following Stallings and Kaskowitz, information about both the presence and use of inventory items is required. While presence of items could be checked at the beginning of the observation day, use was recorded as it occurred. It should be recalled that previous pilot efforts involving over 40 environmental variables had not been encouraging. The Physical Environment inventory contains all the pilot items that showed some potential for discriminating classrooms; it includes clustered seats and room dividers (variables 2-3 in Table 2.1), and in addition itemizes some components of the summary equipment variable (Table 2.1, variable 1), such as projectors, tape players, and the like. To this list were added items from the Stallings and Kaskowitz inventory that seemed appropriate to Alum Rock learning environments (e.g., science equipment and objects, arts and crafts materials, indoor play equipment, magazines, nontraditional reading material, and displays of student work). Finally, items were added that figured prominently in mini-school self-descriptions (e.g., musical instruments, mathematics and reading centers, learning kits, and posted rules and schedules). The resulting Physical Environment inventory includes less than half

the variables investigated in the pilot study, and those it retained were assessed differently. While the pilot instrument recorded the number of each type of item (a source of considerable unreliability), it did not record use (prima facie a more important consideration than presence). Thus, the final version of the inventory was expected to provide a more reliable and relevant account of the physical environment.

Five-Minute Observation

The third part of the instrument, the Five-Minute Observation, represents the greatest expenditure of measurement time and effort, containing the kinds of information deemed most likely on the basis of pilot work to supply the major determinants of classroom diversity (see App. A). Like its counterpart in the Stallings and Kaskowitz instrument, the Five-Minute Observation comprises a set of variables that are repeatedly assessed at five-minute intervals (an interval that also falls within the range of Risley and Cataldo's (1973) PLA-Check sampling frequencies). Four consecutive Five-Minute Observations (a continuous 20-minute period) are used to describe each observed classroom activity, with four activities represented on each occasion of observation.

The Five-Minute Observation begins with a social structure "snapshot" of the classroom, a procedure recommended by Stallings and Kaskowitz (1974), Prescott et al. (1975), Barker (1968), Risley and Cataldo (1973), and other proponents of behavioral observation. The snapshot accounts for all of the students in the class, and provides the data for generating the group size variable from the original factor analysis (item 4, Table 2.1). The diversity instrument does so by dividing them into those working in groups and those not in groups. The students not in groups may be tutored by an adult, a peer, or a cross-grade tutor. Otherwise, they are assumed to be working alone. Then the number of student groups is recorded, followed by the number of students in each group (or, in each of the four most numerous, should the number of groups exceed four; cf. Risley and Cataldo, 1973). Subsequent information in the Five-Minute Observation is recorded for each of these groups.

The first set of variables for which values are supplied relative to each group is called Diversity variables; it includes information about locations, tasks, materials, and mode of work. With respect to each type of information, it is assumed that more than one description may be relevant to each group. For instance, within any group, students may be working on both language and cultural projects employing both texts and locally developed materials. The diversity variables together with grouping patterns represent the information provided in pilot work by factors interpreted as reflecting differentiation and individualization (variables 5-8 in Table 2.1). These factors, it should be recalled, seemed most promising as classroom diversity measures.

Another set of items also assessed for each group, termed Behavioral variables, includes descriptions of ongoing group processes. Many of these variables were drawn from dimensions that the pilot investigation identified as constituting a classroom interaction factor: student-student interaction (Table 2.1, items 35-36), student-teacher interaction (Table 2.1, items 21-32), student participation (Table 2.1, items 38-39), student cooperation or competition (Table 2.1, item 37), and kinds of feedback received by the students (Table 2.1, items 23-26). Other items attempted to index the hypothesized "routinization" factor, which was represented only by the pacing variable in pilot work (Table 2.1, item 10). In the present instrument, self-other pacing is included along with information about task structure (group versus individualized), task leadership (teacher versus student), and the presence of an adult with the task group (Table 2.1, items 18-19). Finally, because bilingual efforts figured so prominently in mini-school self-reports, use of Spanish by groups for instructional interaction was also recorded.

The Five-Minute Observation provides, on a repeated basis, a substantial amount of information about activities in the classroom. In this respect it is similar to the Stallings and Kaskowitz instrument. However, unlike the Follow-Through observation, which focused on the

behavior of individual students and teachers, the observations described here focus on the behavior of groups of students and teachers. In this respect, the Five-Minute Observation is more akin to the RAMOS event form designed by Calfee and Hoover (1974) for observing and recording reading and mathematics groups at work in the classroom.

Classroom Log

The Five-Minute Observations are followed by a section of the instrument referred to as the Classroom Log (see App. A). To be completed once only, after all five-minute observations were finished, they consist of about 50 brief objective statements describing teaching style, teacher classroom management, and student behavior. These items were taken from pilot investigation together with the literature review and mini-school self-reports, and reflect dimensions of substantive interest that could not appropriately be assessed within the constraints of the five-minute observations. On the basis of two or more hours of observation, observers were asked to rate each descriptive item that was characteristic of the classroom on a four-point scale ranging from "never" to "frequently."

With respect to teaching style, some items represented summary-level information directly relevant to five-minute observations (e.g., the classworks by group-interactive processes. The teacher spends his/her time interacting with students). Other statements, however, reflected teaching practices believed to be important in characterizing classrooms and not represented in the five-minute observations (e.g., the teacher makes use of student ideas, the teacher employs direct instructional methods, the teacher's communication focus is individual students). Finally, other items came from mini-school self-reports and had not been successfully piloted (e.g., there is discussion of feelings, behaviors, beliefs, and attitudes, team teaching is employed). Teacher management items, in contrast, had a narrower focus. They sought to establish specific classroom control processes (what recitation patterns are

used, whether written work is assigned and how it is corrected, and methods for giving feedback) and represented structure and management measures in the pilot factor analysis (Table 2.1, items 29-30). Last, student items were centered primarily on two concerns: student mobility, which had shown promise as a classroom discriminator in the pilot study (e.g., students are expected to stay where they are unless given specific permission to move about, students stay in one place); and student choice, another important pilot dimension (e.g., students initiate questions, students are involved in self-instruction, students select their own activities, students engage in self-criticism). The log closes with an unusual events record, a blank area where observers were instructed to note any significant and unusual classroom events that were not adequately represented in any of the prior observational items.

Summary Scales

Ten Summary Scales comprise the last part of the Classroom Observation Sheet (see App. A). These five-point scales require high inference responses to items reflecting very generic dimensions that could underlie a multidimensional model of learning environments. The first five dimensions refer to basic self-descriptive themes of Alum Rock mini-schools. Next are dimensions reflecting two important large-scale constructs arising in pilot work based on the literature review, individualization and differentiation. Finally, two global teacher and student dimensions are employed, tapping teacher task orientation and student involvement. The intent of the Summary Scales was to obtain the judgment of an experienced observer regarding the extent to which factors hypothesized to account for major between-classroom differences are apparent. Completed once at the end of the observation period, the scales sought to index generic dimensions of learning environments on which more molecular observations had been conceptually based.

Teacher Questionnaire

As we have indicated, a Teacher Questionnaire was used to supplement information obtained by observation (see App. A). Questionnaire

items can be most easily organized for discussion purposes in terms of the aims of the instrument. The Teacher Questionnaire first attempted to obtain the teacher's molar-level view of the learning environment. This objective was served primarily by two questions. One of these (question 1), taken from the Stallings and Kaskowitz teacher questionnaire, asks teachers to respond in terms of a five-point rating scale to 13 bipolar dimensions reflecting different teaching philosophies, goals, and procedures (e.g., children work independently versus under adult supervision; emphasis on emotional needs versus subject matter). The second question (question 7), taken from The Rand Corporation's study of reading programs in the Los Angeles Unified School District (Armor et al., 1976), presented 12 items representing organizational or teaching techniques (e.g., individualized instruction, open classroom). Teachers were asked to indicate which if any of these techniques they employed. Both questions had been found useful in discriminating programs. In fact, the question taken from Stallings and Kaskowitz was the only item in their teacher questionnaire that discriminated teachers by Follow-Through model (Stallings and Kaskowitz, 1974).

Another purpose of the Teacher Questionnaire was to solicit kinds of information not readily obtained through observation. It will be recalled that the pilot study indicated that teacher structure and management self-report items might distinguish classrooms. Consequently, the questionnaire sought to determine how teachers controlled the learning process (e.g., ability grouping by subject matter; frequency by subject matter of group change, of assignments, and of testing; type of feedback). In addition, the questionnaire seemed a useful medium through which teachers could elaborate on certain observational items (e.g., specific curricular material used; how often students are referred outside the classroom to special resources). A third kind of information elicited by the questionnaire had to do with total amount of time spent by adults in learning-related activities (e.g., hours per week given to different subject matters; hours per week spent in preparation, in meetings, and the like; hours per week of team-teaching; and hours per week of help from paid aides

and parent volunteers). These questions helped to complete the representation of the classroom derived from observations.

Finally, it was hoped that the Teacher Questionnaire would be useful for studying classrooms in other alternatives systems and for comparing them with Alum Rock classrooms. For this purpose, the questionnaire contained a set of items related to decentralization of decisionmaking. They inquired about the degree of influence in decisions regarding disciplinary procedures, curriculum materials, and lesson content exerted by principals, teachers, parents, and students. In addition, a set of attitudinal items was included. Some items in this set had been examined in the pilot study and had seemed potentially applicable for differentiating teachers (primarily sex role and minority attitude items). Supplementary items were drawn from the teacher questionnaire accompanying an ETS-designed test battery (CIRCUS) for preschool children (Anderson and Messick, 1973). These items reflect attitudes about educational processes (e.g., task orientation, teacher- versus student-centered instruction). The Teacher Questionnaire was administered to all teachers of classrooms involved in the diversity study after the second occasion of observation, and completed the information base.

THE SAMPLING UNIVERSE

In Alum Rock, the mini-school was the instructional entity providing diverse educational options within the schools, and the basic organizational vehicle through which teachers participated in decentralized decisionmaking. Chapter 1 has given an account of the district context within which this demonstration was implemented.

By the fourth year of implementation, 14 of the district's 23 schools (10 elementary schools, 4 middle schools) had chosen to organize mini-schools. The 14 voucher schools generated a total of 50 mini-schools, with an average of 3.6 mini-schools per school. Each mini-school contained from two to 13 classrooms, with an average of five. A voucher school typically involved three mini-school program types. Concurrently, the nonvoucher schools operated on a

more conventional basis in terms of both decisionmaking strategies and instructional organization. However, the average number of classrooms per school building (18) was roughly comparable for nonvoucher and voucher schools. The sampling plan was designed to enable investigation of two related issues: (a) whether empirical clusters of classrooms exist that coincide with mini-schools or groups of mini-schools; and (b) the variance among nonvoucher classrooms relative to the variance among voucher classrooms.

THE SAMPLING PLAN

These issues were most efficiently approached by limiting the sample to the elementary school level (K-5). Elementary school seemed most appropriate to our efforts since its self-contained classes provided a more stable and consistent view of classroom structure and processes than the constantly changing classes of the middle school. However, kindergarten and primary level classes (grades 1-2) were omitted since there were too few of these classes for their exclusive use, and since the kindergarten and primary classroom experience would be sufficiently unique that generalization to the higher elementary grades would be impossible. Special education classes were excluded from the sample for essentially similar reasons. Further, each special education classroom contained a heterogeneous classification of students, suggesting that there was probably no single consistently employed educational approach within them.

Of the remaining grades, a decision was made to sample only 3rd and 5th grade classrooms. Third grade afforded the only feasible grade level for representing younger students, while choosing fifth grade maximized the age spread between the younger and older elementary students. Where the classrooms were multigraded, efforts were made to sample those classrooms with significant proportions of 3rd and 5th grade students. Nevertheless, some 2nd and 6th grade students were inevitably included. The classrooms sampled, then, represented primarily the 3rd and 5th grades.

The voucher schools were structured so that classrooms were nested within mini-schools, which in turn were nested within schools. To minimize confounding, it was necessary that each voucher school be represented by two different mini-schools, each of which was represented by at least two classrooms. Within the appropriate grade level, then, voucher mini-schools were exhaustively sampled. For each of 27 mini-schools, two classrooms, representing the 3rd and 5th grade levels, were selected. The voucher sample consisted of 54 classrooms.

The sample of nonvoucher classrooms was drawn from the nine nonvoucher elementary schools. Two 3rd grade and two 5th grade classes were randomly chosen from each school, for a total of 36 nonvoucher classrooms. Thus, the total number of classrooms selected for the diversity study was 90. Using the same sampling criteria, 16 additional classrooms were identified as substitutes if any of the original sample were unable to participate. Classrooms in the diversity study sample were to be observed twice, by a single observer on each occasion of observation, with order of classrooms observed counterbalanced over the two occasions. A detailed technical account of the sampling plan appears as App. C.

This sample differed slightly from the sample that was fielded. Through a clerical error, some voucher classrooms that did not conform to sampling plan requirements were added to the sample. Thus, the fielded sample was identical to the theoretical sample in terms of the total number of classrooms samples (90), and in terms of the number of voucher (54) and nonvoucher classrooms (36), but a discrepancy occurred among voucher classrooms that affected the distribution of classrooms within mini-schools. While the proposed sample called for at least two appropriately graded classrooms in each of 28 mini-schools, the fielded sample in fact consisted of 20 mini-schools that contained two classrooms each, three mini-schools that contained three classrooms each, and five mini-schools that contained only one classroom in the appropriate grade range. This discrepancy between the theoretical plan and the fielded plan resulted in a reduction in the sample of voucher classrooms that could be used for data analysis. The final analysis sample consisted of 74 classrooms, 40 in the voucher demonstration.

A final small pilot study was undertaken to permit development of observer training and data collection procedures, and to allow inter-observer reliability estimates. For this purpose, a sample of six classrooms was selected from six mini-schools housed in six different schools. Observers were fielded in pairs so that each classroom was seen by four observers on two separate occasions of observation. Order of observation of the six classrooms was counterbalanced for the two occasions of observation, and each observer was paired with every other observer on some occasion.

DATA COLLECTION

Data collection efforts were designed to avoid both the specific problems that arose during pilot classroom observation studies and the more general problems of observation in the natural setting. Perhaps the most critical concern was to obtain data from an appropriate sample of classrooms. Since the preceding year's pilot classroom observation study was a voluntary effort, teachers felt free to cancel their scheduled visitation. The result was a small sample of observed classrooms unevenly distributed across mini-schools and schools. The confounding caused by unbalanced samples resulted in the loss of up to 50 percent of the mini-schools and 67 percent of the schools during data analysis. A major concern in planning data collection strategies, then, was the risk that classroom cancellations (or nonparticipation for other reasons) would significantly alter the planned sample.

In accord with this concern, a data management system was developed to minimize the risk of cancellation and maximize the probability of appropriate substitutions or rescheduling should cancellations occur. A complete set of information regarding the sample classrooms was housed in the site office in a three-way filing system. Records were organized on the basis of teacher identification number, and were cross-filed both by school and by classroom sequence number (the number assigned to each classroom in the sample representing its position in the study). A critical aspect of data management was a log system involving three subparts: A Telephone Log recorded scheduling contacts with teachers as well as requests for information and other telephone transactions with school personnel; a Confirmation Chart recorded communications

regarding scheduling or rescheduling of sample classroom (ensuring that each observation appointment was reconfirmed the previous day and that the teacher and students would be present at the designated time); a Master Data Log followed each observation record from its original identification and assignment to its shipment to the Rand Santa Monica Office; finally, a Special Attention File reported events requiring immediate action (e.g., a research assistant reporting ill, a teacher unexpectedly absent on an observation day, cancellation of a class, and the like). Items in the Special Attention File were flagged until appropriate action was taken and its nature entered on the Master Data Log.

Another major data collection concern involved the fielding of observers. Previous classroom observation efforts had been flawed by the uneven distribution of observers over occasions, resulting in widely different amounts of experience in observing. Our concern, then, was to ensure an even distribution of observers over occasion of observations, and to ensure that observers were randomly distributed for an even spread of observer error. These concerns were addressed through a structured system of observer assignment to classrooms. First, to minimize observer cancellations that would alter the entire observer/classroom schedule, observations were scheduled only four days per week, with the fifth day free for the observers' personal business. This schedule also allowed time for updating the data management system, and for confirmation of the next week's scheduled classrooms. To control for experience, observers were assigned approximately the same number of classrooms, and the interval between classroom visitations was made comparable. Finally, observers were assigned approximately equal numbers of voucher and nonvoucher classrooms, and efforts were made to keep them blind to classroom condition. Since the observers had worked as substitute teachers in the Alum Rock school district and in surrounding school districts, many of them were familiar with the voucher and nonvoucher schools. Thus, we first determined the degree of familiarity of each observer with each teacher/school in the sample. We then avoided assigning observers to schools with which they had prior contact, and cautioned

the observers against discussing the demonstration status, or other classroom/school characteristics, with the other observers. While these precautions did not ensure that observers were blind to condition, they did increase the likelihood that observers would be unfamiliar with demonstration status of particular classrooms.

A third major issue affecting data collection was the climate of opinion toward the classroom observation study (see Chap. 1). Although participation in the study was mandated by NIE, we were concerned that some subset of the sample would attempt to avoid observation by taking a sick day, scheduling a field trip, or other evasive moves. Further, we were concerned that the teachers' severe apprehension over evaluation would result in atypical, or even "staged" classroom behavior. Thus, important considerations during data collection were to minimize observer obtrusiveness and establish teacher confidence in the study. These difficult problems were confronted through direct contacts with teachers and other district personnel, as well as through the training of observers to promote sensitive interaction with the teachers. Contacts with practitioners are described in more detail below.

PRELIMINARY MEETINGS WITH DISTRICT PERSONNEL

NIE's fourth-year contract with the district required participation in the classroom observation study. However, the climate of suspicion that permeated the district because of budget cuts and pending personnel reductions made it necessary to meet with the Alum Rock staff to address their concerns about the purposes of the study, the study procedures, the required participation, and the ultimate uses of the data. During early December 1975, Rand researchers held a series of meetings with voucher principals and teachers, voucher mini-school coordinators, nonvoucher principals and teachers, representatives of teacher organizations, and administrative personnel.

The Rand researchers attempted to alleviate apprehension about the content of the observation instrument by providing examples of the types of observational categories included in the study. Observation categories that were considered too subjective or otherwise objectionable

were either revised or eliminated. Most concerns about the observational process were laid to rest by providing detailed information about the number of occasions of observation, average length of observation, criteria for selection of classroom observers, advance notification procedures, and the like.

The overriding concern about confidentiality and the potential use of observation data by district officials was somewhat mitigated by the establishment of a site office separate from the district's Research and Evaluation Department.[1] Qualms about the confidentiality of the research were further allayed by the fact that neither school name nor mini-school name appeared on the observation instrument. A teacher's name appeared on the instrument only on the identification label, which was to be torn off and presented to the teacher at the end of the observation period.

Finally, the teachers were given the name of a Rand staff member to call (collect) if any questions or problems arose during the implementation of the observation study. All meetings concluded with Rand assurances that the observations would be conducted as unobtrusively as possible.

PILOT STUDY

The pilot study was conducted from January 6 to January 23, 1976. Its purpose was threefold: to determine whether the preliminary instrument generated reliable data, to determine the field reliabilities of the observers, and to determine the field efficacy of the data management procedures.

Observer Selection and Training

In response to teacher concerns about the use of district substitute teachers as classroom observers, efforts were made to recruit a pool of observers who had prior experience in classroom observation studies unrelated to the Alum Rock study. Because of the time requirements for the Alum Rock study, however, no previously trained observers

[1] During the 1973 and 1974 pilot observation efforts, school district facilities were used.

were available. Concerted efforts were then made to obtain and train substitute teachers from neighboring districts as observers.

Selection of the observers was made by a Rand staff member assigned to the Alum Rock site office. Descriptions of the job were placed in the newspaper and posted in school district personnel offices. Twelve applicants who responded prior to the scheduled pilot training session were interviewed. Selection criteria included previous teaching experience with children comparable to the Alum Rock sample in age, ethnicity, and income level, and availability to work four days weekly throughout the pilot and the major studies. On the basis of these criteria, five white female substitute teachers who ranged in age from 22 to 28 years were hired. Four of the teachers were to serve as regularly scheduled classroom observers, and the fifth was to serve as an alternate.

Training for the pilot study was conducted January 6-9, 1976. Pilot training involved a combination of seminar instruction and field practicum work using a final draft of the observation instrument. Seminar instruction focused on the operational definitions of terms included in the observation instrument and specific procedures to be used while conducting the observations, as well as the nature and importance of interobserver reliability. Slides and videotapes of Alum Rock classrooms featuring instructional materials and classroom processes were used to facilitate pilot training. After two days of seminar instruction, observers conducted trial observations in 2nd and 6th grade classrooms (which were not part of the study sample). Field reliabilities were continuously calculated during the trial observations.

Pilot Data Collection. The pilot study design involved six 3rd and 5th grade classrooms drawn from different mini-schools located in six different voucher schools. As we have noted, each observer was paired with every other observer, and each classroom was observed twice by a different pair of observers. Table 3.1 depicts the planned assignment of observers to classrooms over both occasions of observation; numbers represent individual observers.

Table 3.1

CLASSROOM OBSERVER ASSIGNMENT

Classroom	Observation 1	Observation 2
1	1, 2	3, 4
2	1, 3	2, 4
3	1, 4	2, 3
4	2, 3	1, 4
5	2, 4	1, 3
6	3, 4	1, 2

A week before the beginning of pilot data collection, letters of notification were sent to the classroom teachers in the pilot sample, with copies to their mini-school coordinators and school principals. The letters explained the general purposes and procedures of the observations; they included a statement of confidentiality and the scheduled observation dates. The letters were followed by confirmation calls to the classroom teachers on the day prior to the scheduled visitation. During data collection, observers reported to the Rand site office approximately an hour before the scheduled observations and obtained classroom assignments from the study staff. Observers then proceeded to the reception office of the assigned school and asked to be directed to the scheduled classrooms.

Once in the classrooms, observers introduced themselves to the teachers, made their assurances of unobtrusiveness and confidentiality, and offered to let the teachers review the data collection instrument at the end of the observation period. Observations were primarily limited to the morning hours, and extended into the afternoon only when one or more of the required reading, mathematics, and other academic or nonacademic courses could not be observed during the morning. On completion of their observations, the observers, while still on site, reviewed the instruments for errors and omissions.

Tear-off labels were then given to the teachers, and the observers left the classrooms no later than 2:30 p.m. At the site office, data collection forms were again checked for error or omissions by the site office supervisor, and observers were debriefed by the Rand training staff on a daily basis.

After pilot observations had been completed, participating teachers were also debriefed. The teacher debriefings, which focused on adequacy of notification procedures, level of observer obtrusiveness, and confidentiality procedures, provided information relevant to the revision of the observation and data management procedures. Debriefing input from the observers also served as a source of revision for these procedures as well as for the draft observation instrument.

Interobserver Reliability. Subsequent to the pilot observations, interobserver reliabilities were calculated as the ratio below for each pair of observers:

$$\frac{\text{No. of agreements}}{\text{No. of agreements + No. of disagreements}} \times 100$$

Calculations were made for each section of the instrument except the classroom roster, which contained teacher self-report information. Obtained interobserver reliabilities are listed in Table 3.2.

Implications of the Pilot Study. The pilot study generated considerable information relevant to the implementation of the larger diversity study. This information served as the basis for revisions in instrumentation, data collection and data management procedures, and observer training. These aspects of the study are discussed below.

First, it was apparent from both the observers' debriefing comments and the obtained interobserver reliabilities that the Classroom Log and Classroom Rating sections of the observation instrument generated data of limited reliability. Since items in these sections required some evaluative judgments, revisions involved clarification in the wording of certain items as well as deletion of other items that were difficult to clarify. Finally, more precise operational definitions of the most subjective items were generated to be used during subsequent training.

Table 3.2

INTEROBSERVER AGREEMENT FOR PILOT STUDY

	Pair Number	Agreement (%)					Total Observations
		Physical Environment	Five-Minute Observations	Classroom Log	Ratings	Overall Ratio	
1.	3,4	88	98	76	90	97	2215
2.	2,4	85	91	64	40	88	761
3.	2,4	94	96	70	60	95	1363
4.	2,3	97	98	62	70	96	1318
5.	1,3	85	95	76	80	94	1828
6.	1,4	88	97	70	90	96	1420
7.	1,5[a]	97	98	78	100	97	978
8.	1,3	88	98	72	79	96	1135
9.	1,2	77	96	70	70	95	1828
10.	2,3	91	97	76	100	97	1715
11.	3,4	100	98	86	100	97	1080
12.	1,4	100	98	76	90	97	1148
	Average	91	97	73	81	95	[1399]

[a]Denotes an alternate observer.

In response to observer debriefing comments, revisions were also made in the Classroom Roster. This portion of the instrument described classroom schedules and other information about students and teachers (e.g., the number of students, their grade level and seating assignments; the number of teachers and other adults involved in classroom instruction) that would provide a context for the interpretation of other classroom data. Initially, Classroom Roster information was to be observed. However, since this information could be more reliably obtained from the teachers, a procedural change was made so that observers asked this information of teachers before the start of class. The Roster was also modified to reflect the variety of scheduling and structural features found across classrooms. For instance, the instrument was revised to reflect the fact that seating and group assignments were fluid rather than static, and that students might be out of class for staggered reading in both the morning and the afternoon.

Since interobserver reliabilities were consistently high, minimal changes were made on the Physical Environment portion of the instrument. Based on the training experience and debriefing comments, however, definitions were rewritten to distinguish math and reading centers; to distinguish games and toys from math manipulatives and puzzles; and to distinguish magazines and other "nontraditional" reading matter from traditional reading matter.

The Five-Minute Observation portion of the instrument, which recorded ongoing classroom activity and organization, also underwent minimum revisions. Specifically, the instrument was modified to accommodate the dissolution of student groups and to reflect more varied tasks and materials used by students as well as additional modes of teaching. Finally, the Behavioral Variables section of the Observation was modified to reflect that at any given moment, students/teachers could engage in both behavioral alternatives (cooperation and competition) simultaneously, or could engage in neither. In summary, the major focus of instrument revisions was to better accommodate the potential variability in classroom structure and processes.

Second, with respect to training and data management, teacher debriefings emphasized the fact that well-informed and politically

sensitive observers could readily allay teacher fears and apprehensions regarding the observation. Thus, a decision was made in observer training to include a discussion of the attitudinal climate in which the study was being conducted, strategies for putting teachers at ease, and strategies for responding to difficult questions or situations. The teacher debriefings also indicated a need for additional scheduling/notification procedures. Consequently, in addition to the initial letter of notification and the followup call, a Reconfirmation Note was developed. This note was to be hand-delivered to the school on the day prior to the scheduled observation.

A final data management issue involved reviewing the data forms for omissions and errors at the site office. On the basis of the pilot study, it was anticipated that the office secretary would have considerable difficulty editing the forms of 10 to 15 observers during the same period when he was making final confirmation calls to the schools the afternoon prior to their next scheduled observations. A suggested alternative that observers exchange and check each other's forms was rejected. It was finally decided that the office manager should split his half-time commitment between fielding the observers in the morning and reviewing the forms in the afternoon. This solution was an uneasy one but seemed the most appropriate given the options available. Budgetary constraints required the termination of his services entirely midway through the second observation.[2]

The Diversity Study

The diversity study was conducted from February 17 to March 22, 1976. During this period, each of the 54 voucher and 36 nonvoucher classrooms was observed twice with order of observation counterbalanced. The 90 classrooms were observed by 14 observers, five of whom were carryovers from the pilot study.

Observer Selection and Training. Using the selection criteria and procedures identical to those of the pilot study, 12 additional observers (11 white and one black) were hired for the diversity study. They ranged in age from 22 to 35 years. All were female and worked in the district as substitute teachers. The twelve newly hired observers, together with the five pilot observers, participated in a

[2]This was an administrative decision over which the researchers had no control; it seriously vitiated the usefulness of second-round data (see p. 74).

five-day training session. The pilot observers, who were knowledgeable about operational definitions and observational procedures, frequently assisted during training by leading small group discussions, by role-playing teacher-student interactions, and by clarifying definitional issues that seemed unique to the Alum Rock classrooms.

A basic training resource was the Training Manual, developed to ensure consistency of training and field procedures across a large group of observers. The Manual familiarized observers with the general purposes of the voucher demonstration and provided information regarding the role of classroom observation in the larger research effort. It also established field appropriate behavior and professional standards, and suggested responses to questions posed by district personnel about the research. In addition, the Manual described the data management system in detail, including specific tasks to be completed prior and subsequent to data collection.

Most important, the Training Manual presented detailed information about the procedures to be followed while making observations. Operational definitions were developed for all terms in the classroom observation instrument. These definitions were constructed after intensive briefings with the pilot observers, and established concrete guides for applying conceptually generated descriptors to ongoing classroom events. These definitions appear as App. B. Because the Training Manual was given to them as a reading assignment the week prior to training, observers were familiar with the observation instrument when they reported to the first training session.

Like the pilot study training, the diversity study training consisted of both seminar instruction and field practicum work. The basic assumption of training was that observers should receive as much field experience as possible as soon as possible. Since the Manual had been reviewed prior to the onset of training, only the first day was spent entirely in seminar instruction, while remaining days were divided between seminars and field experiences. The first day's seminar instruction was designed to promote conceptual understanding of the observation instrument and of the objects and events to be observed. To this end, each section of the observation instrument was discussed separately, with videotapes, slides, and actual

objects (e.g., traditional versus nontraditional reading material, packaged learning kits, and the like) used where appropriate for illustrative purposes.

The field practice sessions constituted a critical aspect of training, since they provided the observers the opportunity to interpret the observation instrument in the classroom context. These sessions, which were conducted in nonsample classrooms, began on the second day and continued mornings for the duration of the training. During each field session, groups of three or four observers (each group including both experienced pilot observers and novice observers) were dispatched to classrooms where they practiced their introductory and exit statements, applied conceptual definitions to classroom items and events, and developed facility in the use of the observation instrument. On their return to the site office, debriefing sessions were held during which each group discussed problems in definition, interpretation, appropriate behavior, and anything else of concern. When disagreements arose that could not readily be resolved in small-group discussion, the issues were presented to the entire group. Rotating small-group membership to avoid the development of idiosyncratic perceptions among subsets of observers, we followed this basic format throughout training. Interobserver reliabilities were continuously calculated and averaged in the upper 80's by the latter training sessions.

Observer trainees were closely monitored, and on the final day of training 14 of the 17 trainees were selected as regular observers. Selection criteria included operational knowledge of observation categories, facility with the observation instrument, and high level of interobserver agreement. The remaining three observer trainees were designated as alternates, and were on call throughout the data collection period. However, individual observers were sufficiently dependable so that only one alternate was needed, on a single occasion, to substitute for a regularly scheduled observer.

Data Collection. Data collection began February 24 and continued through March 22. Although March 22-26 was reserved for rescheduling, all rescheduled classrooms had been observed by March 22. Observations

were made in 90 3rd and 5th grade classrooms on two occasions of observation by a single (different) observer each time. Table 3.3 displays the distribution of observed voucher classrooms by mini-school and mini-school type.

Preparation for observation and actual observational procedures were similar to those used during the pilot study, but were somewhat more elaborate. A Master Classroom Observation Schedule listing observer assignments and classrooms to be observed for each day was the basis of confirmation procedures. Using the Master Classroom Observation Schedule, letters of notification were sent to teachers, mini-school coordinators, and principals a week prior to the scheduled observation. These letters were followed by a brief reconfirmation note a few days prior to the scheduled observation. On the day before the scheduled observation, final confirmation calls were made to the classroom teachers. Confirmations (or cancellations) of scheduled observations were noted on the confirmation chart, which was then checked against the master schedule. If a class was cancelled or was not confirmed, it was noted in the Special Attention File for rescheduling, and the scheduled observer was notified of the change in observation dates. For confirmed classrooms, the Site Director matched identification numbers of pre-labeled observation instruments with an identical set of information on pre-labeled Research Assistant Assignment Check forms (RAAC, the daily assignment sheet for each observer). Both documents were then checked against the classroom identification numbers listed for that day on the Master Schedule. The RAAC forms and the instruments were then compiled for use by the observers the following day.

On the scheduled observation day, the observers reported to the site office about an hour prior to the scheduled classroom visit, signed in on the Daily Visit Log, and obtained their classroom assignments--on RAACs--from the Site Supervisor. Before departing for their scheduled classrooms, observers compared the information on the instrument with the information on the RAAC.

Each observer arrived at the assigned school at least 20 minutes before the class was scheduled to begin, introduced herself to the

Table 3.3

DESIGN CHART FOR OBSERVED VOUCHER CLASSROOMS

Classroom (N=54)	Mini-School (N=28)	School (N=10)
	Basic Skills	
1*	Action in Learning	A
1,2	Tradition Plus	B
1,2	Success School	C
1*,2*,3*	READ	E
1,2	Total Experience	F
1,2	Little Schoolhouse	G
1*	BEST	H
1*,2*	Basic Skills	I
1,2,3	Basic Academic	J
	Individualized	
1*	Learning for Life	A
1,2	Communication Plus	D
1*	Learning Unlimited	E
1,2	Small World	F
1,2	ILP	H
1,2,3	Continuous Progress Learning	H
1,2	BRL Silas	I
1*	School 2000	I
1,2	ILP	J
	Open Structure	
1*,2*	Learn by Doing	A
1,2	Daily Living	B
1,2	Adventures in Learning	F
1,2	People to People	G
	Multicultural	
1*,2*	Cultural Arts	B
1,2	Bilingual/Bicultural	C
1,2	Bilingual/Bicultural	G
1,2	Multicultural	J
	Fine Arts	
1,2	Fine Arts	I
1,2	World of Fine Arts	D

*Denotes classrooms deleted from analytic sample because of nonreplication, inappropriate actual grade range, special education status, or other innapropriate classifying properties.

school secretary, and asked to be directed to the scheduled classroom. In the classroom, the observer introduced herself to the teacher, assured the teacher of her personal unobtrusiveness and of the confidentiality of the study, and offered the teacher the opportunity to review the data collection instrument after the observation was complete. The observer then attempted to obtain the Classroom Roster information from the teacher. If the teacher was unable to provide the information before class began, the observer obtained it during the first recess. Following the initial exchange with the teacher, the observer located an unobtrusive seat from which she could easily view the classroom. When it was necessary to move around the classroom to make more accurate observations, the observer did so as quietly and inconspicuously as possible.

The Physical Environment Inventory, the simplest, most objective section of the instrument, merely required the observer to note the presence and use of any of the listed items. The Five-Minute Observations (see the description of instrumentation above) were considerably more complex. This section of the instrument contained four 20-minute sequences, each consisting of four five-minute observation periods. For each 20-minute sequence, the research assistant observed a reading period, a math period, one other academic period (science, history, social studies, and the like), and one nonacademic period (excluding recess or physical education). Twenty-minute sequences were started after the scheduled activity had begun so that transition behavior (changing seats, getting supplies, waiting in line, and so on) would not be observed. Such sequences were separated by 10-minute breaks.

After completing the Five-Minute Observations, observers filled out the Classroom Log and the Summary Scales, based on their earlier judgments about the teacher's classroom organization and management style. Observations were typically finished during the morning hours and extended to the afternoon only when one or more of the required periods could not be observed in the morning. After the total observation instrument was completed, the observer, while still in the classroom, reviewed the form for errors and omissions. Just before leaving, the observer gave the tear-off label from the instrument to

the teacher and allowed the teacher to examine the instrument. On the observer's return to the site office, she again signed in on the Daily Visit Log, and noted on the RAAC any comments about unique aspects of the classroom or the teacher. Observers were debriefed and observation instruments were edited by the site supervisor before the observers signed out.

Although the data management strategies were fairly onerous to implement, they successfully minimized teacher resistance to the study, cancellation of classes, and observer cancellations. Of the 90 classrooms observed, five had to be replaced with substitute classrooms. Of these five classrooms, two teachers were taking extended medical leave (one was pregnant), one classroom dissolved, and two teachers expected to have half or fewer of their students available during the four weeks of the study. The efficacy of the strategies employed was also evident in the low level of observer cancellations. Only one observer cancelled a scheduled observation. Further, observers were fairly equally distributed over observations, with some 85 percent of the observers visiting approximately equal numbers of classes. Finally, the lack of hostility and resistance from teachers after the first round of observations suggested that our approach to them had been successful.

Observer debriefing at the conclusion of data collection yielded interesting comments about both the instrument and what they observed. With respect to instrumentation, observers continued to have problems with the Classroom Roster throughout the diversity study. Because of the rather fluid movement of students in and out of the classroom, the total number of students present in the class or the total number of students sent out of the class during the observation day for an academic reason frequently exceeded the number of students enrolled. Similarly, the number of different types of adults in the classroom frequently exceeded the total number of adults, since teachers often counted a single adult in several different categories (e.g., a bilingual parent volunteer).

Observers remarked that on the whole, the classrooms seemed quite well organized. Teachers typically adhered to the schedules they gave

the observers for the Classroom Roster, and seemed quite similar in their structuring of classroom activities. For instance, virtually all classrooms were run in an orderly fashion, with most activity being teacher initiated. Seats, groups, and tasks were typically assigned by teachers, and there was little integration of work and play within the classroom.

In terms of equipment and materials, observers reported few math or reading centers. In contrast, science centers were numerous, although they were used very infrequently. Very few teacher-developed or student-developed materials were present in the classroom, and the only nontraditional reading material mentioned was National Geographic magazine. The primary distinction among classrooms seemed to be one of quantity of materials. Classrooms which observers reported as having considerable quantities of programmed materials, audiovisual aids, learning games, and the like, tended to be voucher rather than nonvoucher classrooms.

Observers reported little discussion, little boardwork, many tests, and a great deal of question/answer drill. Little interaction was observed, they believed, because students were busy with assignments and assignments tended to be carried out in fairly traditional fashion. Little use of Spanish as an instructional medium was observed. Classrooms emphasizing basic skills appeared to differ from other classrooms by providing mainly individual work without much variety either in terms of tasks or teaching modes. However, observers thought they saw fairly high levels of student participation across classrooms. The level of task interruption was very low, and groups seldom dissolved.

In general, despite their self-descriptions, most voucher (as well as nonvoucher) classrooms appeared to be oriented toward basic skills. Few instances of individualized, bilingual, or learn-by-doing classrooms were in evidence, and fine arts (as an educational orientation) was virtually nonexistent. Some innovative classrooms were observed: an integrated curriculum, a performance-contracting classroom, and a program uniquely designed around centers. Generally speaking, observers were unable to distinguish voucher from nonvoucher classrooms on the basis of classroom structure, content, or process.

Chapter 4

ANALYSIS

As we have noted, the goal of data analysis was to build and interpret a spatial model of the variation among educational alternatives in Alum Rock on the classroom level. This task was begun with two assumptions. First, it was assumed that if there is genuine diversity of choice (e.g., if there are differences among classrooms), this diversity should be visible. What should be visible is behavior, and perhaps organizational structure. Consequently, the model of educational variation should build on a core set of low inference variables that index these attributes. Second, it was assumed that a "traditional" classroom type exists and is recognizable, structurally and behaviorally. Such a classroom, we assumed, is likely to be like those experienced by most Americans: the teacher-talk-student-listen model, where work is separate from play and is done quietly in one's assigned seat. If these are the features that make the traditional classroom recognizable, the search for diversity should look for variations on and departures from these features. This line of reasoning also led to the conclusion that the analysis should depend primarily on reasonably direct indicators of structural and behavioral diversity.

With these guiding assumptions, a core set of variables was selected to form the basis of the descriptive space within which observed classrooms would be located. The components of the DIVERSITY variables appear under that heading on pages 4-5 and succeeding pages of the observation instrument (see App. A), together with a fifth grouping variable. The latter is a function of the number of observable groups and the number of students in each group, established at the beginning of every five-minute observation. These variables, as the literature review suggested, are generally taken as important descriptors of educational programs. Moreover, it appeared to us that the basis of concurrent diversity in these variables is spatial/structural differentiation--that cross-sectional variability in tasks, for example, required such differentiation. In short, we saw diversity in these variables as

necessary (but not entirely sufficient) conditions for departures from the traditional classroom in the direction of greater variety. This same argument would suggest that observations of greater classroom variety would be registered across the set of diversity variables. It might, therefore, be adequate to measure only one instead of all five. But given the chances of error and the fallibility of any particular observation, it seemed best to retain five basic diversity variables.

From this viewpoint, other classroom phenomena were more appropriately regarded as consequences of the spatial/structural diversity assessed by the DIVERSITY variables. For example, diversity in BEHAVIORAL variables (see items under that heading in the observation instrument, pp. 5ff., App. A, and operational definitions in App. B) seemed relatively unlikely in the absence of structural variation. (To anticipate, our analytic results do provide empirical support for using the DIVERSITY variables as the basis of our constructed descriptive space. As we will discuss in more detail below, fitting theoretically relevant interpretive variables into this space yields results which tend to validate our choices. Moreover, measures of the extent to which the descriptive configuration based upon the DIVERSITY variables is congruent with that based upon BEHAVIORAL variables suggest a good degree of commonality.)

Finally, choosing to build the model of educational variation primarily on the DIVERSITY variables meant that self-descriptive data obtained from teachers would be used mainly for interpretive purposes. In fact, there is some evidence of congruence between observational and self-descriptive spaces for this set of classrooms. Suppose that, from a multidimensional scaling of the same set of objects based on derived distances from two distinct sets of variables, there are generated two distinct sets of coordinates for the objects in a space of K dimensions. Further, suppose that one is interested in knowing the degree of congruence between the configurations. That is, given an N x K matrix (of objects by coordinates), and an N x M matrix derived from independent data ($M \leq K$), one might wish to know the degree to which the spaces were congruent. One method of assessing the degree of congruence is to fix the first matrix as a "target" matrix, and apply to the second an

orthogonal transformation which has the effect of rotating it toward the target to the position of best fit in the least squares sense.

To assess the degree of congruence between two-dimensional configurations based on most of the observational variables and a large selection of self-descriptor variables measured in this study, we applied a program called PROCROT (see App. E for a further description), using the observational configuration as the target. Among the indices of congruence provided by this program is a matrix of correlations between the columns of the target and the rotated matrix. In this case, the correlation between the respective first factors was .50, and between the second factors, .12. The first is fairly respectable, but the second is actually lower than the off-diagonal correlations. As we will show, one plausible reason for this amount of lack of fit is that the dimensionality of the scaling solution is too low for the observational data. For present purposes, however, the point is simply that the two descriptive spaces are almost surely related.

DESCRIPTION AND DEFINITION OF DIVERSITY VARIABLES

We considered spatial/structural differentiation to be a necessary condition for behavioral diversity, and we sought to differentiate classrooms in terms of within-classroom diversity. (We did not predetermine the dimensionality of the descriptive space. We simply allowed the DIVERSITY variables to form its basis.) To assess spatial/structural differentiation, we evaluated five variables. The first four variables were evaluated simply by counting the number of different values the variable took concurrently (i.e., within one five-minute observation) in the same classroom:

NLOC = the number of different locations in active use by at least one student group during a single five-minute period.

NTASK = the number of different types of defined activities (e.g., reading, music), that were occurring simultaneously in the same classroom.

NMAT = the number of different materials in simultaneous use in a classroom (e.g., textbooks, slide projectors).

NMODE = the number of different modes of learning/teaching which occurred simultaneously in the classroom (e.g., boardwork, discussion).

All of these variables aim to assess the diversity of activity in a classroom at a given time, and each represents a degree of departure on a single dimension of activity from our assumed traditional classroom. Conceptually, they are coherent, which is why we treat them as indicators of a single underlying concept. On the other hand, they need not be empirically correlated, although they are expected to be.

The fifth variable, which we call the lumpiness variable and name LUMP, has a more complex derivation but a straightforward intent. It originated in a search for a single measure to describe the observable distribution of students in a classroom, taking into account both the number of discrete groups and the number of students in each. Number of groups, it seems, is a prima facie indicator of diversity. Unevenness of distribution of students across groups (where the smallest "group" might be a single student) seems a reasonable indicator of allowance for divergent interests, preferred modes of behavior or learning, and so on, measured at the individual student level. Our intent was to avoid having the measure depend directly on the absolute number of students in the classroom, since this is a variable only very partially under the control of individual teachers or mini-schools. Hence, we sought a measure that would allow some weight for absolute number, but would depend more heavily on distributional variables.

A complete discussion of the properties and construction of the lumpiness variable is not presented here, since it appears elsewhere (Barker and Bikson, in preparation). We will simply suggest its obvious properties and what it does. Briefly, it orders classrooms directly in terms of the number of groups in the classroom. For a given number of groups, it orders them directly in terms of the unevenness of the distribution of students in the groups. It also assigns a slightly higher value in any of these classifications to smaller classrooms. Finally, the value for the traditional classroom (where the entire class constitutes one group), is zero. Hence, the more heterogeneous the distribution of students, the higher the value of LUMP.

Technically, LUMP is defined as the absolute value of the sum of the natural logarithms of the proportion of students (by classroom) in each group. Its properties depend upon the well-known properties of positive integers that

$$(n - i)^k \geq (n - j)^k , \qquad i < j \leq n$$

and

$$(n - i)^k \leq (n - i)^m , \qquad m > k, \; i \leq n.$$

Logarithms are used in order to avoid very small fractions, and the absolute value to order classrooms directly instead of inversely. This variable is perhaps the clearest indicator of spatial/structural diversity within a classroom. It is most directly related to NLOC, and conceptually (though not necessarily empirically) related to the other DIVERSITY variables. These five variables taken together provide the basis for the classroom descriptive space.

Characterization of the Classroom

In our description of the data gathering method, we mentioned that each classroom was observed on two different days, by a different observer each time, and that the same central set of variables was scored on 16 separate occasions each day in four consecutive 20-minute intervals. Theoretically, this procedure would have been adopted *a priori* to attempt to cover the contingencies of both stability and change, in the ANOVA sense of separating variability within from variability between classrooms. Further, we believed that classrooms could in fact be characterized as reasonably enduring stylistic entities, rather than as randomly fluctuating series or (except perhaps on a lengthy time scale) as evolutionary processes. Hence, the primary purpose for repeated sampling of behavior within and between occasions was to achieve greater stability for our measures. If we were, in fact, dealing with relatively enduring characteristics, within-classroom departures from the characteristic value would tend to be small relative to between classroom differences assuming that classrooms differ. In any event, given a classroom characteristic value, aggregation across numerous samples would tend to give greater stability to the estimate.

The clearest chance for finding a confound in this estimate lay in the possibility that real differences within a classroom might occur with subject matter differences, so that aggregation over all time samples might yield an uninterpretable and certainly noncharacteristic value. This naturally suggested the possibility of separate aggregation and analysis for different subject matter sessions, or at least subject matter groups. Under the field conditions of this study, such a procedure would have caused difficulties, primarily because of considerable nonuniformity of observational coverage, which in turn resulted from differing distributions in and over time of subject matter sessions (insofar as these were empirically separable). In any event, this approach was regarded as unnecessary, since, if one assumes that clear and basic stylistic differences among classrooms do exist, it seems reasonable to suppose that these differences would carry across various subjects. In addition, one reasonable difference in style might itself be the tendency not to make marked distinctions between traditionally differentiated subject matters. If these considerations turned out to be wrong, then it would still be reasonable to aggregate over all time points, yielding an intermediate value on the affected variables. And if we were correct, the problem would not arise.

As it happened, the problem did not arise save in one case, which was revealed by the main analysis. Our debriefing of pilot observers suggested strongly that such stylistic differences as would appear would tend to be uniform in time, which seemed to vindicate our a priori approach. In short, we expected both theoretically and, following pilot work, empirically to find characteristic classroom differences, insofar as they existed, to be uniform within classrooms over time samples. Thus, since distributions of variables tended to be reasonably symmetric if not thoroughly normal, or else decidedly skewed as a result of genuine outliers, it seemed quite appropriate to work with classroom means in the principal analyses.

CONSTRUCTION OF DESCRIPTIVE SPACES

Spatial representation enables people to perceive patterns in ways that columns of numbers do not. In addition, as the preceding section

shows, spatial/structural differentiation is the key to our conception of the phenomena under study. We set out to describe classrooms on the basis of observational data in a DIVERSITY space. For that purpose, we adopted an approach called multidimensional scaling (MDS), one variant of which permits the construction of descriptive spaces without making metric assumptions about the distances on which the construction is based (see Barker et al., 1975, for a discussion of the basic MDS principles involved).

The MDS approach adopted here rests on the idea that one may make sense of a set of data based upon the observation of discrete objects by considering it as a matrix of measures of the proximity of the objects to each other. Numerous examples of this may be cited. Perhaps the most common social science example is a matrix of product-moment correlation coefficients. (While one usually thinks of correlations among variables, this is more from custom than necessity. Practitioners of what is sometimes called profile analysis would, in this context, treat correlations among objects.) The notion of proximity here is, however, quite general. Shepard (1972) gives as a (non-exhaustive) set of rough synonyms ". . . similarity, substitutability, affinity, association, interaction, correlation . . ." In short, the construct of proximity includes almost anything such that one may reasonably say that the measure says something about how close objects are to each other. Now, if we are able to judge how close objects are, we are also judging how far apart they are, or their distances from each other. The notions are complementary. In what follows (adhering to common usage), we will refer almost exclusively to distances, but one need only recall that a smaller distance should imply a greater degree of proximity, and conversely.

Distance, however, is not a notion to be taken absolutely seriously. We assume, given the data usually available, that about the best claim we will be able to make is that object a is farther from object b than it is from object c. In short, we will take seriously only the ordinal properties of the data. We will be willing to bet on greater than, but not on how much greater. On the other hand, this caution is probably not as restrictive as it seems. The present profusion of MDS

techniques and analyses may be said to have originated from Shepard's (1962 a, b) observation that, from purely ordinal information, one may generally recover interval or ratio information. Using Monte Carlo techniques he showed that, for any useful number of objects (say, 10 or so), given only the rank order of their interpoint distances, one could recover (up to a constant of proportionality) the interpoint distances themselves with nearly perfect fidelity, if one had been restricted to preservation of the rank order. This is easy to visualize, for instance, in a one-space: given, say, 10 objects to be arranged in the space so that the rank order of the 45 (or $n(n - 1)/2$) interpoint distances is preserved, one rapidly finds that there really is only one way to do it. If one begins to shift points around, one is almost bound to violate that rank order. In this way, then, from information about the rank order one has, in fact, recovered the interpoint distances themselves up to a constant of proportionality, which we may think of as a unit of measurement.

One obvious problem for MDS is that one is almost required to know the dimensionality of the space before starting, and this is something that one generally does not know when confronted with a proximity matrix. This difficulty does not have an analytic solution. What is wanted, though, is not a perfect fit of the spatial model to the object, but one that is good enough. This idea brings us to a discussion of what MDS algorithms do. Perhaps the simplest basic description is Shepard's (1972):

> We seek simply that configuration of n points in the (Euclidean) space of smallest possible dimensions such that, to an acceptable degree of approximation, the resulting interpoint distances d_{ij} are monotonically related to the given proximity data in the sense that
>
> $$d_{ij} < d_{kl} \text{ whenever } s_{ij} > s_{kl}.$$

This description implies that there is some measure of departure from perfect fit, some loss function which can be evaluated. A rather

extensive discussion of the class of loss functions generally used in MDS algorithms, generically called STRESS, may be found in Young (1973). STRESS is a variance-like function which may be usefully regarded as an error sum of squares or mean squares. When the fit is perfect, STRESS equals zero. It is further useful to note that in computing STRESS, the algorithm will, given a configuration of points, compute the d_{ij} and compare them with the given s_{ij} (which may or may not be distance estimates in the sense given by Shepard above). Note in particular that no special form of function relating the d_{ij} and s_{ij} is postulated. The requirement of monotonicity, essentially equivalent to a requirement of rank order preservation, is all, and is relatively quite loose. To make the point more sharply, recall that most factor models postulate that the function which relates measure and factor(s) is strictly linear, a most restrictive assumption relative to that of monotonicity.

If the comparison of d_{ij} with s_{ij} made in terms of STRESS yields a value greater than zero, the algorithm will figuratively move the points around a little, reestimate the d_{ij}, again evaluate STRESS, and so on. This process reiterates until, as nearly as the algorithm can judge, for a given dimensionality, no further improvement is possible. (Clearly we are skipping a number of concerns, including questions of local minima, rational starting configurations, and so on. We do so in the interest of clarifying the overall analysis approach.) The phrase "for a given dimensionality" is crucial. Given an N x N matrix of Euclidean distances, we know that we can fit them perfectly in an (N - 1)-space. We may also be able to fit them perfectly in an (N - k)-space, where $N \geq k > 1$, or at least obtain a "good enough" fit, meaning a fit in a space such that STRESS is close enough to zero. One of our aims is parsimony. We would, other things being equal, much prefer a two-space to a three-space, and either to anything greater. Also, we cannot claim that our data are "true," i.e., error free, in the sense that they equal population values. These considerations suggest that "good enough" is not a phrase that invites a precise answer.

Let us consider what "good enough" might reasonably mean in the context of the well-known D^2 measure. As we know,

$$D^2_{ik} = \sum_{j=1}^{J} (S_{ij} - S_{kj})^2.$$

Suppose that none of the points vary at all on some dimension, which we may call the $(n - 1)^{th}$. Clearly, it would make no contribution to D^2, and we would simply exclude it. Now suppose the points vary somewhat but very little relative to their variation on other dimensions. Then it is clear that the squared term of the sum computed from that dimension would be uniformly small relative to the others, and could be omitted without changing any of the D^2_{ik} very much. If that were the case, the fit in $(n - 2)$-space would be, for us, good enough. While the STRESS measure is not perfect, it does depend to a large extent upon a numerical evaluation of the relative contributions of the dimensions in which it is computed. A solution for us, then, will be good enough if (1) STRESS is not unduly large (say, $\leq .10$); (2) there is a reasonable number of dimensions; and (3) we obtain an interpretable configuration. Of course, point (3) is something about which no algorithm can provide any guarantees.

In summary, given a matrix of measures of proximity among a set of objects, we attempt to fit them into a space of as few dimensions as possible such that violations of a perfect monotonic relationship between distances estimated from that space and the data themselves are as few and small as possible. The particular program which we use to do this is called POLYCON (Young, 1973), and is further described in App. E. Our principal reasons for choosing this particular program are practical. It is a relatively recent program, which means it takes advantage of past work on algorithms of this kind. Its user language is simple and flexible, so that it is relatively easy to alter limiting parameters (primarily, the number of objects which can be scaled--this study involves an unusually large number of objects). It incorporates a variety of measurement models. It has been extensively tested, and it is readily available.

DISTANCES

Not too long ago, space was what was described by Euclid: a place where we lived, and the area we had to work with. Recently, other

kinds of spaces have been explored. One critical distinction about a space is afforded by how a point is permitted to move in it. This defines what is called the metric of the space. If measures of distances are defined or computed, the metric says what is permissible. For example, in what is usually called the city block metric, operational distance is not how the crow flies, but the length of the path taken by the usual pedestrian. In contrast, Euclidean space implies an operational "crow flight" distance. For research tasks which involve constructing a spatial model by beginning with a proximity matrix, the nature of the "real" space is not immediately revealed. In such a case, one might take as a criterion the spatial construct that seemed most parsimonious and interpretable. In the present instance, however, we were confronted not by a proximity matrix, but by a profile matrix from which proximities or distances were to be computed (a discussion of derived versus direct distances is provided below). Hence, the kind of space with which we would be dealing would depend on the manner in which we chose to define the distances between object profiles.

We thought of variations of classrooms in a space in Euclidean terms, and we have done our substantive analytic work in a space of this kind. We did, however, attempt some preliminary scaling in the city block metric and found the results less parsimonious in terms of both dimensionality and interpretability. In addition, again as a preliminary step, we explored a variant of what Young (1973) calls a multiplicative or FACTOR model, which amounted to a principal components (PC) analysis of the data profiles. (The product matrix was, of course, of deficient rank, but that was not a concern.) Our conclusions based on both of these preliminary analyses were principally two: (1) that greater dimensional parsimony would be provided by a Euclidean space analysis than by the others, especially in comparison with the PC model (a not uncommon finding, given the more restrictive assumptions of the PC model); and (2) that the general form of the configuration would not be materially altered by choice of model, since it was in broad outline quite visible in each of our approaches. Without more exhaustive exploration, we opted for a Euclidean distance analysis with dimensionality to be determined from the data. That is to be assumed in all scaling results reported below.

For any distance metric, the principal working output of an MDS routine is a set of plots showing the relative locations of the units scaled in the obtained space. As would be expected, for N dimensions we usually obtain N(N - 1)/2 plots representing each distinct pair of dimensions. These are the vehicles for attempting to answer the main question: Why are the points located where they are? More specifically, we might observe that certain points are close to each other while others are far apart, or that there appear to be points defining opposing poles of some dimensions, or that we seem to see relatively self-contained clumps or clusters of points with relatively wide distances between them. Some or all of these patterns might result, and it is this resulting configuration that we seek to understand.

Some objects carry more intuitive meaning with them than others. A matrix of Morse code signal confusions, for example, might exhibit two dimensions when scaled, and we might be able to explain the observed configuration by noting that one dimension of variation appeared to be a function of the length of the signal, the other of the number of dots (or dashes) the signal contains. A scaling of judged similarities between pairs of U.S. Senators might appear to anchor one dimension with liberalism, and one with conservatism (Shepard, 1962 a,b). Our classroom units, on the other hand, would seem to carry little or no apparent meaning with them since, apart from the data, we know nothing about them. That is, we cannot interpret a configuration of classrooms simply by looking at the names of the scaled points. Rather, for determining the basic dimensions that govern perceptions of the classrooms--to map as nearly as possible the total perceptual space--our mode of procedure is not ideal. Perhaps these considerations will be clearer in terms of the notion of undirected estimation of similarity.

A typical situation for collecting undirected estimates of similarity (or differences) would involve presentation of all possible pairs of objects to judges. Judges would be directed to indicate, on a suitable scale, their estimate of how similar (or different) the objects are. These judgments, suitably aggregated or considered separately, would constitute appropriate data for MDS analysis. It should be noted that the judges are not given any direction regarding what attributes

of the objects to consider in arriving at their judgments. The point would be to see what determinable dimensions seemed to underlie these estimates, i.e., to determine as far as possible the attributes to which the judges attended in making their judgements. This would amount, then, to obtaining similarity estimates, scaling them, and seeking to understand what the judges might have had in mind.

On the other hand, we can conceive of judgements made under totally opposite conditions, in which the attributes were specified precisely. For example, instead of asking judges to estimate general similarity among a set of automobiles, we might ask them to estimate their similarity in terms of horsepower or sportiness. This method might be chosen either because it was just these attributes in which we were interested (in which case, with respect to horsepower, we would have an external check), or because we thought that when similarity estimates were made these were the attributes (more likely, the attribute) that accounted for the undirected judgments, and we wanted to check our notions against data. In the present case, it should be clear that even though we are not using judges' estimates of similarity (or distance), the situation is much more like the second scenario than the first. That is, we perforce assume that we have included the important characteristics of classrooms in our study, and if we were asking judges to estimate similarities, we would direct them to attend to these attributes (at least, we would specify diversity). In fact, we would have liked to obtain, in addition to the observational data we collected, such undirected estimates of similarity. That is, we would have liked to know whether judges appear to attend to the same things when they make their estimates; if not, whether we can identify ways in which they differ; and, in either case, to what they do seem to attend. One obvious problem with this procedure is the difficulty of conceptually specifying the nature of a classroom. A more practical difficulty is the matter of familiarizing a set of judges with a sufficiently large number of classrooms to permit them to make similarity judgments about all pairs. Lacking similarity estimates, the present study depends entirely on derived distances for developing its spatial model, and for interpreting the resulting configuration.

Finally, then, we come to a consideration of the (dis)similarity or distance measures actually employed in this study. Basically, they are computed according to formula (1) above. That is, for each pair of classrooms, the difference between their values on each of the five diversity variables is computed, the result squared, the squares summed, and the square root taken. This produces a square symmetric matrix with all main diagonal entries equal to zero. The entries in this matrix also meet the standard metric axioms for distance (Shepard, 1972). One further point should be noted. We have adjusted for differences in scale, since the D_{ij} are scale free, but the component variables are not all on the same scale. If the scales themselves are very different, the scales with the largest ranges (which could be simply a matter of arbitrarily different units) would naturally contribute the largest components to the distance measures. If the variances were very different, the same thing would happen. Since, at the outset at least, we had no reason to favor one variable over another, we sought to neutralize such scale differences. This is done simply by scaling each variable by its own standard deviation, so that all variables have unit variance.

We should end the discussion of distances by underscoring the difference between a (dis)similarity matrix derived in this way from essentially profile data, and one obtained directly. It is true that with this procedure the metric axioms are automatically satisfied, but this is not conceptually as important as the fact that the scaling configuration can only be based upon the information used to construct the input matrix, i.e., on the DIVERSITY variables defined above. In that sense, the (dis)similarity estimates are completely directed, and we can have little confidence that the resulting space coincides with any particular perceptual space. We necessarily depend upon our belief that we have taken into account the components of classroom diversity, but we have no means, from our data, of checking it. Thus the representational validity of the obtained spatial configuration depends in large measure on the adequacy of the set of variables specified for observation.

ANALYTIC SAMPLES

A primary research interest guiding the analysis was the question of whether empirical clusters of classrooms could be found roughly to coincide with mini-schools or groups of mini-schools. Secondary interest was directed at the amount of variation to be found among nonvoucher classrooms relative to that among voucher classrooms, since mini-schools did not exist in nonvoucher schools. It was also considered possible that, in a combined sample, certain types of classrooms, perhaps in part coincident with mini-schools but common to both voucher and nonvoucher schools, might appear. To determine whether empirical clustering was a result of factors associated with schools per se, with mini-schools, or with individual classrooms, a fully replicated sample of classrooms was required (see App. C). The organization of educational service delivery in the voucher demonstration was doubly hierarchical. In the design sense, classrooms are nested within mini-schools nested within schools, which produces some inevitable confounding. It was thus essential that each voucher school be represented by at least two mini-schools, each of which must in turn be represented in the sample by at least two classrooms.

Since exhaustive sampling was feasible in the grade range for which the study was designed, the voucher classroom sample was drawn in an attempt to represent each of the 28 mini-schools covering this range by two classrooms. This process did not yield full within-school replication, but this was a minor concern. A more vital question was, in design terminology, the apportionment of variation among mini-schools relative to the between-classroom residual. As shown in detail in App. C, failure to attend to this replication requirement in the 1975 pilot study resulted in a loss of 67 percent of the schools and 50 percent of the mini-schools sampled. Unbalanced sampling (due to voluntary participation of classrooms in that investigation) resulted in a large amount of confounding of units, which rendered them useless for purposes of this kind of analysis. The present sampling plan was to avoid such loss of resources. However, field-generated schedule changes resulted in a reduction in the sample of replicated voucher classrooms from 54 to 40, a loss of about 25 percent.[1] This sample of 40 classrooms, then, was the principal analytic sample for the study. It is sometimes referred to as sample V, and

[1]This was an administrative decision over which the researchers had no control.

its distribution over mini-schools, mini-schools types and schools is given in Table 3.3.

Although each classroom in the sample was observed on two separate days, only data from the first observation are included here. Field editing problems with the second round of observations due to the loss of supervisory staff noted in Chapter 3 resulted in data that were not as reliable or complete as the data from the first round. These second round data were ultimately omitted from the analysis.

The second analytic sample, usually named NV, includes 34 non-voucher classrooms, distributed as explained in the procedures chapter. The original sample of 36 was reduced by the failure of two teachers in the sample to provide self-report data. In summary, the primary analytic sample is V, with 40 classrooms. After presenting analyses for this sample, we will present analyses in which the two analytic samples, V and NV, are combined. The nonvoucher classrooms were not analyzed separately since, for reasons which will become clear in the discussion, such analysis appeared to be unnecessary.

BASIC SCALING RESULTS

As explained above, the scaling of classrooms (in contrast with the interpretation of resulting configurations) was for analytic purposes based on Euclidean distances derived from the five DIVERSITY variables. The spatial configurations of the units (classrooms), then, are based on just these distances. The first question to be answered was that of dimensionality. Initially, we scaled the voucher classrooms (the sample, N = 40, is that called V) in 3, 2, and 1 dimensions simply to get an idea of the proper dimensionality (as defined above). Since it was immediately clear that one dimension would not do (for example, obtained STRESS = .30, and departures from good fit were numerous and striking), we concentrated on the other two potential solutions. It was also clear, and confirmed by analysis, that no more than three dimensions were necessary. The question thus became to determine whether a two-dimensional solution was good enough, and we concluded that it was not. The improvement in obtained STRESS (Young, 1973) in moving from two to three dimensions was relatively substantial. The two-space yielded a final STRESS = .06, which for this measure is not bad but is borderline. However, for the three-space we

found that STRESS < .02, which is as close to zero as any anlysis is likely to come. Moreover, as we briefly discussed above, the rotation in two-space of a scaling solution based on self-descriptive variables to one based on observational variables produced a relatively poor fit. We concluded that the problem came from inadequate dimensionality. This conclusion is further supported by the comparison of rotations in two- and three-space of solutions based on BEHAVIORAL variables to those based on DIVERSITY variables.

In the two-space case, the correlations between columns of the DIVERSITY-based and BEHAVIORAL-based configurations after rotation were .60 and .08. Again, while the correlation between the first pair of columns or dimensions is respectable, that between the second pair is essentially zero. Taken alone, these results would seem to suggest at least the existence of a very strong first dimension, and perhaps that the additional dimensions are unimportant. This latter inference, however, is controverted by the STRESS of the respective solutions. (The first result is, however, strongly supported by all of our analyses, including the principal components solution mentioned earlier.) Perhaps, then, taken alone, the rotation results might suggest a common first dimension and essentially independent second ones. But in fact, taken alone, the two-space rotational results are misleading. An analogous rotation in three-space of DIVERSITY and BEHAVIORAL configurations yields comparable correlations between dimension pairs of .67, .42, and .54. These values not only suggest that the difficulty with the two-space analyses stemmed from inadequate dimensionality, but also that the third dimension is quite significant, at least analytically. (Unlike MDS solutions, rotations are analytic.)

Finally, we note that, as we might expect, the second dimension of the two-space solution seems to represent an unhappy confounding of what are, in fact, different dimensions. Not only does a three-space solution present compelling evidence for the usefulness of the third dimension, it considerably clarifies the second. To illustrate these conclusions, we present in Fig. 4.1 a perspective drawing of the three-space configuration where points represent classrooms in sample V. Perhaps the most striking feature of this configuration is its essential definition in terms of only three classrooms, those numbered 1, 2, and 3. More particularly, we see a swarm of 37 classrooms distributed fairly evenly

Fig. 4.1 — V classrooms in the 3-space configuration

around the origin of coordinates without any compelling definition, and three classrooms relatively distant from the swarm as well as from each other. To underscore this point, as well as to illustrate the striking change in going from a two- to a three-space solution, Table 4.1 presents the coordinates in these respective spaces of the three classrooms mentioned.

Table 4.1

COORDINATES OF DEFINING CLASSROOMS IN 2- AND 3-SPACE

	Classroom		
Dimension	1	2	3
	2-Space[a]		
1	0.83	2.11	3.10
2	0.84	1.13	0.24
	3-Space[b]		
1	0.72	1.26	3.36
2	0.84	2.50	0.98
3	1.21	0.38	0.39

[a] STRESS = .06.

[b] STRESS = .02.

The clarification that results from going from two- to three-space is evident from the data in Table 4.1. Note first that the two-space analysis suggests that we are dealing only with one strong dimension, and that classrooms 2 and 3 principally define it. This picture changes radically when we consider the data for these classroom units in three-space. It now becomes clear that unit two is primarily varying on the second dimension (note that this unit alone changes very much in its coordinate on the first dimension), and that unit one (which formerly was seen to vary relatively little on any dimension), now seems to define the third dimension. In fact, it is evident that the strong

dimensional definition apparent in this configuration derives almost entirely from three classrooms, each of which seems to define a separate dimension.

This is a striking and unexpected result. It suggests strongly that if we consider the sample classrooms in terms of the DIVERSITY variables, 93 percent (37 of 40) of them are essentially indistinguishable, while the remaining seven percent (3 of 40) are quite distinctive from the rest and from each other. It further suggests that a clustering approach to interpretation of the configuration would be futile. We could reasonably expect four clusters, one with 37 members and three more with one member each. Evidently, if we are to understand these results, we must attempt to interpret the dimensions themselves.

The Outlier Problem

Before proceeding, however, we must try to deal with a question that has probably occurred to the reader. In the language of linear models, we are confronted with an outlier problem. Ordinarily in experimentation, the immediate attitude toward outliers is suspicion (see, for example, Draper and Smith, 1966; or Cochran, 1970). The first step is to determine whether the outliers are simply the result of error or artifact. If that determination is made, the outliers are generally excluded from the analysis on the grounds that they apparently belong to a different (and scantily represented) population from the one of immediate interest. In this case, our initial attitude was similar. In particular, we wished to assure ourselves that we were not looking at results which were artifactual or representative of some sort of measurement error.

Accordingly, when confronted with these results, we began a careful investigation of the protocols from which the data for these three classrooms were extracted. The outcome of this study was reassuring. Overall assessment, as well as detailed examination, showed that those classrooms were extremely consistent at each occasion of observation as well as within observations. That is, the data were consistent and clearly not erroneous. Examination of the observational detail was supplemented by an investigation of observers' "unusual events" logs and

a check of debriefing records made at the end of the observations. These investigations confirmed the conclusions based on the more detailed data. Observer comments about these classrooms on the instruments unanimously indicated that they were seeing something out of the ordinary, while debriefing notes confirmed that these outliers had been noticed and remarked as unusual by observers present at the session. Hence, we concluded that the outliers were real.

At this point, however, our attitude diverged from the experimental. Conceptually, all of these classrooms were members of the population of interest. To exclude some because they were different would undermine the basis of the entire study. Recalling our initial hypothesis that we would be able to detect interesting differences (outliers) if they were there, it was clear that our research goal was to discover empirically different populations. In short, our aim was to assess heterogeneity, not to suppress or exclude it. The fact that the heterogeneity found was not in just the form we expected did not invalidate either the goal or the methods.

The issue of expected heterogeneity of classrooms, however, introduces important questions about population and samples. To this point, while acknowledging that the classrooms in our analytic sample constitute neither the population of voucher classrooms, nor--more to the point--the population of mini-schools within the grade range of interest, we have discussed the results as though this were not true. In fact, the first qualification is not necessarily a serious one. Because of the small size of many mini-schools and their tendency to employ multigrade classrooms, our observational sample does include the bulk of the eligible units. The second qualification is more serious. Recalling the discussion of the composition of the V sample, 25 percent of the observed classrooms were eliminated because they were not members of fully replicated sets. It should be emphasized that, since one of our basic aims was to determine the extent to which empirical groupings of units seemed to be explained by mini-school or school membership, as against simple classroom differences or similarities, it was essential, whatever the specific analytic strategy we undertook, to exclude unreplicated units from the analytic sample. (This procedure and its rationale are set out at greater

length on pp. 157ff., App. C.) Had the three outliers not been members of fully replicated sets, they too would have been excluded from V. In that case, we would surely have drawn very different conclusions about diversity in Alum Rock classrooms.

In any event, it would seem that, short of observing all of the classrooms in the grade range of interest, we have in fact included a set of units as thoroughly representative as possible. Yet it is fair to say that this is the result as much of good luck as of thorough planning. Moreover, given the results we got, we really should conclude that any amount of sampling is suboptimal, if not simply dangerous. Perhaps the problem may most easily be clarified by recalling the nature of the current research undertaking. A leading reason for sampling is the desire to estimate population parameters without undue effort, but this rationale has little relevance for us. We had no reason to expect that we were dealing with some empirically homogeneous population at all--just the reverse. In short, the present research was not construed as an attempt to sample a homogeneous population; rather, we were presented with a conceptually homogeneous population whose empirical homogeneity or heterogeneity was just the object of interest. From this point of view, sampling is irrelevant.

Finally, the outlier problem raised questions about scaling solutions. We noted earlier that, in order to avoid having our scaling solutions driven by unit or range difference, we scaled each variable in the DIVERSITY set so that all variables had unit variance. This procedure assumes, however, that the scaling factors either equal or are unbiased estimates of the population values; but, as should be obvious, the values obtained for the standard deviations are very different depending upon whether we include or exclude the three outliers. We have argued that exclusion of the outliers--whether deliberately for cause, or inadvertently by sampling--would defeat the purposes of our research. However, it may nonetheless be of some use to consider whether the differing scaling factors make much empirical difference.

The most accessible approach to this question seemed to be to compare the configuration of the 37 homogeneous classrooms obtained by simply eliminating the three outliers after obtaining the configuration

with a configuration obtained by eliminating outliers first and scaling the remaining 37 units alone. Hence, we rotated one such configuration to the other, with the following results: The correlations between the three pairs of dimensions from the two solutions were, in order, .96, .85, and .82. It is evident that employing the very different scaling factors made very little difference in configuring the 37 common and homogeneous classrooms. When we consider that irrespective of the scaling factor these classrooms are homogeneous, so that even small changes in data values which might plausibly be attributed to measurement error are likely to result in some changes in rank order, this result is all the more impressive.

Classroom Similarity Within Mini-School

We had set out to determine whether or not similarities among classrooms were a function of mini-school membership. The short answer to this question, based upon analyses undertaken here, is negative. The evidence for this conclusion appears in more detail in Figs. 4.2, 4.3, and 4.4, where geometric symbols represent classrooms in the sample. These figures present the three two-space projections of the three-space configuration that appears in Fig. 4.2, and include two complementary kinds of information. The lines connect the two (and sometime three) classrooms in the V sample which are members of the same mini-school. Two conclusions may immediately be drawn: (1) none of the three outliers are members of the same mini-school; and (2) classrooms in the same mini-school do not cluster together. Said differently, a classroom is about as likely to be nearest a classroom from a different mini-school (more likely, in fact), as it is to be nearest one from the same mini-school. In short, whatever may be influencing this configuration, it is not mini-school membership.

The geometric symbols in each figure convey the typology of mini-schools devised by the evaluation department of the Alum Rock district, based largely upon nominal self-report data from the mini-schools as described in Chap. 2. As we have noted, the typology is a mixture of subject emphasis (e.g., FINE ARTS), and self-reported process (e.g., LEARN BY DOING). Since the configuration we present may be said to be

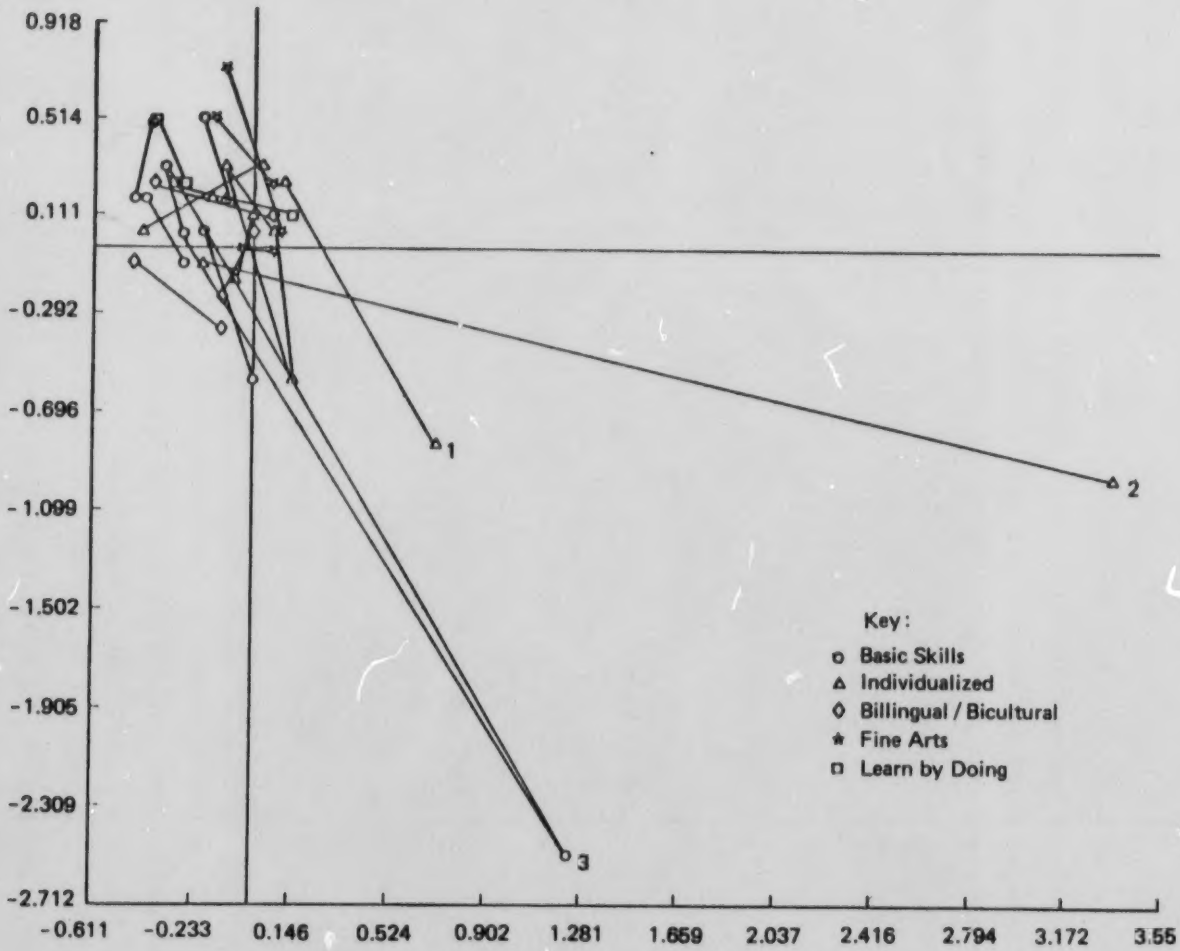

Fig. 4.2 — V classrooms and mini-school connections:
Dimension I x Dimension II

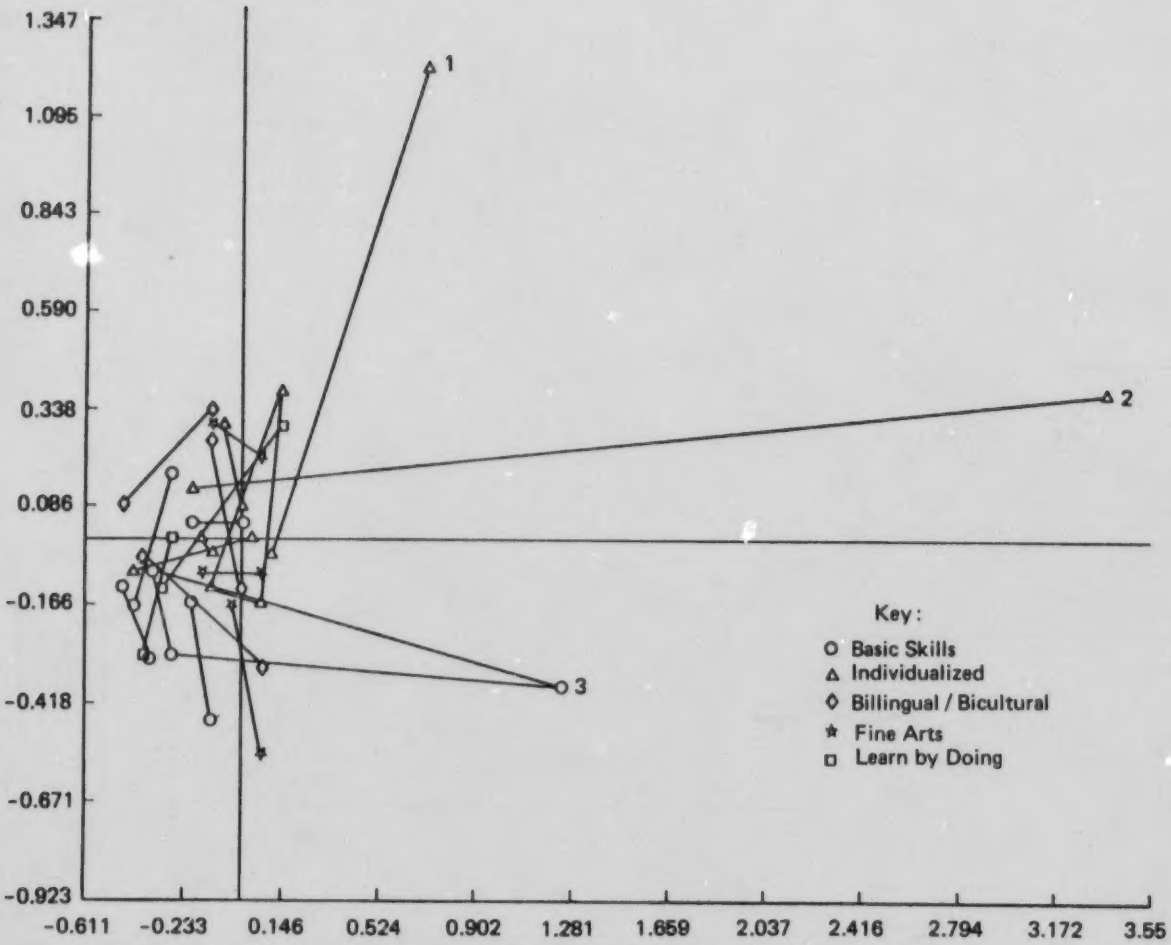

Fig. 4.3 — V classrooms and mini - school connections:
Dimension I x Dimension III

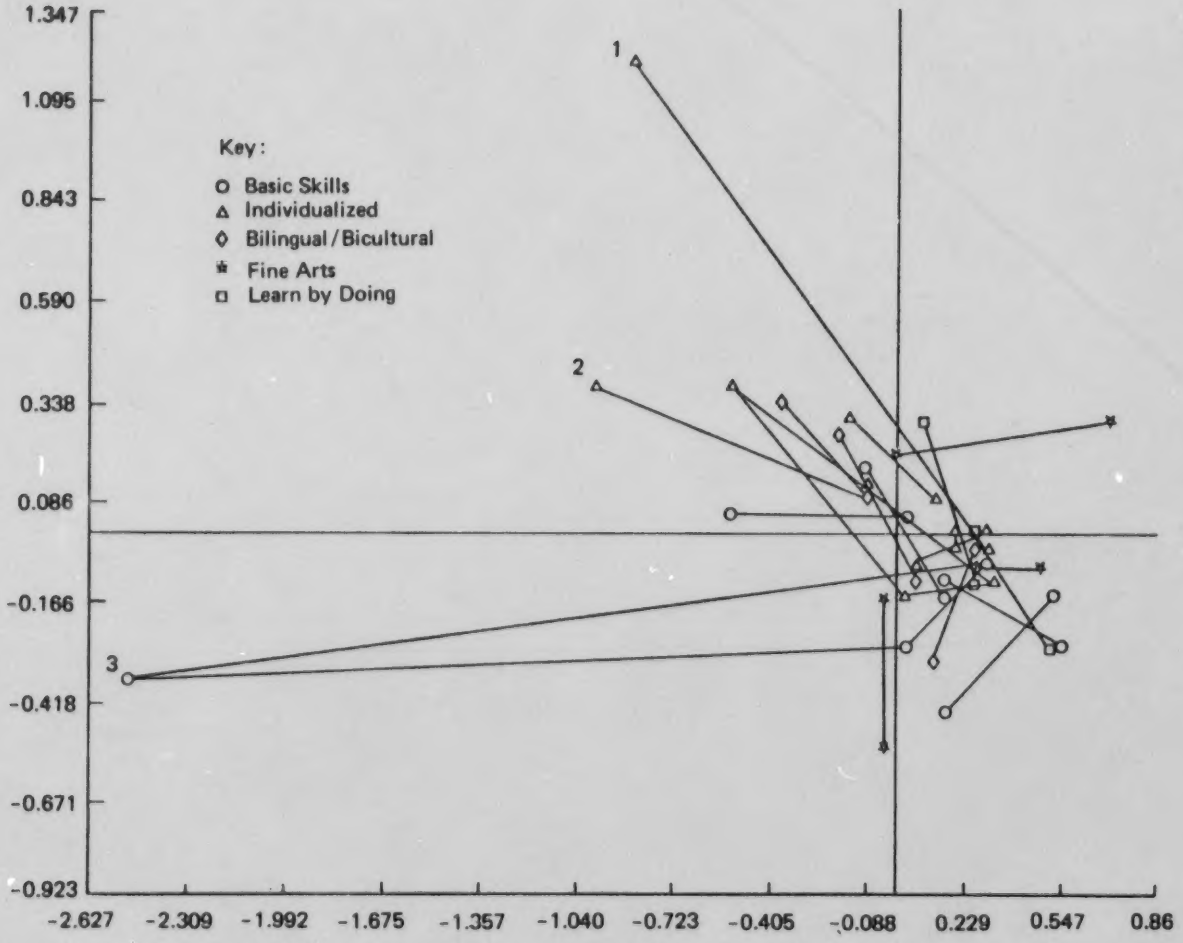

Fig. 4.4 — V classrooms and mini-school connections:
Dimension II x Dimension III

based largely if not entirely upon process variables, it may be of some interest to determine the congruence between observational and Alum Rock nominal groupings. As is evident from the figures, no spatial separation is apparent. Units show no tendency to separate and cluster by geometric symbol. Of the three outliers two are of the same type (INDIVIDUALIZED), a not surprising outcome given the basis of the configuration. This result might be taken to indicate some degree of reasonableness in the Alum Rock typology. On the other hand, one of the outliers is called BASIC SKILLS, and the basis for this is not apparent. Perhaps more to the point, most of the INDIVIDUALIZED mini-schools are part of the clump around the origin of the coordinates--which suggests that, if the DIVERSITY variables in fact capture process individualization, the bulk of the INDIVIDUALIZED programs do not reflect their ideology in practice. On the whole, then, we are led to conclude that neither mini-school membership nor announced congruence of program goals appears to drive process similarity. However, this is negative information. In the next chapter, we attempt to understand something about what <u>does</u> appear to explain the obtained configuration.

Chapter 5

INTERPRETATION

In Chap. 4 we discussed the difficulties of interpreting configurations of objects that do not carry with them some prima facie definition. This set of classrooms is an example of a configuration of essentially meaningless objects. It is not clear why the empirical outliers are outliers, except in the sense that relative to most other classrooms their values on the DIVERSITY variables are different. We have seen, for example, that mini-school types do not imply observed differences even when, as in the case of INDIVIDUALIZED, we might expect this to be true. Clearly, then, an interpretive strategy is required. The one we have adopted is to introduce independent information which is theoretically relevant. We have already noted that, given the obtained configuration, we cannot learn much from a cluster approach to interpretation, which suggests that the major dimensions themselves are the most likely vehicles for arriving at an explanation. Unfortunately, as we have seen, these dimensions seem to be anchored largely by single classrooms and not, as we might wish, by groups of classrooms whose characteristics are sufficiently similar to suggest some rationale for their configuration.

An alternative course would be to find one or more variables whose scales coincided with the observed dimensions, and which were independent of (that is, not used in the construction of) the observed configuration. In practice, interpretation could be undertaken with something less than that. A set of vectors in the space, with the outliers differentiated from the bulk of the classrooms (and from each other) in meaningful ways, would provide a usable foundation.

Fortunately, an algorithm for constructing such vectors is available and is realized in a program called PROFIT (see App. E). Briefly, this program requires as inputs the coordinates of some scaling configuration and values for the scaled objects on some other variable(s), called "properties" in the program. By regressing the property values on the coordinates, PROFIT maximizes the correlation between the

property values and the projections of the object points on a vector in the space. For example, if the values for the classrooms on some variable such as pacing ordered the classrooms in the same way as their projections on the first dimension of the configuration, we would find a correlation very near 1.00, and a vector essentially coincident with that dimension. That degree of congruence would be unlikely, but we could find some variables for which this state is approximated. Of course, if we looked long enough, we would be likely to find some variable (Rubin, 1974) that would produce a correlation of 1.00, but we would also be likely to find that the result was nonsense. If we are to make sense of the configuration, it is necessary that the variables chosen for examination in this way also make sense. That is, they must be "theoretically relevant," where the "theory" in question is that given by the conceptual framework for selecting the DIVERSITY and other variables for study.

The ten variables chosen for purposes of interpretation may be understood as follows. The first, called PACING, is simply a low inference direct observational variable (see the observation instrument, App. A, p. 115). The observer, for each group observed, was to judge directly whether the group was working at a pace imposed by someone else or at a self-imposed or self-chosen pace (see App. B for the operational definition of this variable). It was assumed that the more self-pacing observed, the more diverse (in our sense) the classroom is likely to be, and the values of the variable follow that expectation.

The second variable, called SPANISH, is defined by the number of instances in which Spanish was used in classroom instruction, broadly construed (see the observation instrument, App. A, p. 115, and the operational definition in App. B). This was included both because of the ostensible emphasis in some mini-schools on bilingualism (note that one type included in the Alum Rock typology is BILINGUAL/BICULTURAL), and because use of first-language instruction would seem to indicate a certain kind of individualization relevant to diversity. This variable, of course, is very low inference.

Six of the other eight variables are scales constructed from a series of overall judgments made by observers as the last observational tasks (see the observation instrument, App. A, Classroom Log and Summary Scales). Having completed all sequences of five-minute observations in a classroom, the observers were asked to make judgments such as "The teacher's questions are open-ended" on either four- or five-point scales. These judgments (or items, in the usual scaling vocabulary) were combined into higher-order scales by an iterative process involving rational judgment (items which seemed conceptually to belong together) and empirical analysis (primarily alpha maximization; see, for example, Bentler, 1968; Armor, 1972). Details of the scale contents are presented in App. D. To summarize, we list the variable names and sample items below. The first two represent teacher activities, the next two represent student activities, and the last two are derived from generic judgments about the classroom as a whole.

Variable name	Sample items
TRADITIONAL	The teacher uses traditional instructional methods; the teacher's focus is not individual students.
ACADEMIC	The teacher provides academic feedback; the class works quietly, individually.
STUDENT-ACTIVE	Students initiate questions, discussion; students are involved in self-instruction.
STUDENT-PASSIVE	Students are in the respondent role; students are quiet unless called on.
DIVERSITY ORIENTED	Emphasis on self-teaching by discovery; emphasis on self-other awareness.
BASIC ORIENTED	Emphasis on basic skill learning; the teacher expects a lot of work to get done.

Finally, we constructed two analogous scales from a series of teacher self-report judgments, primarily about their classroom practices and orientations (see the teacher questionnaire, App. A, items 1a-1m). The resulting variables and sample items were as follows (see App. D for a more complete account):

Variable name	Sample items
SUBJECT CENTERED	Students work under adult supervision; emphasis is on subject matter.
STUDENT CENTERED	Students gather information on their own; students initiate interactions with adults.

When the ten variables described above are fitted to the configuration displayed in Fig. 5.1, they appear as vectors (in contrast with the classroom points) in the space. That is, they have both known orientations to the configuration coordinates and directions in the space, so that they can be pictorially represented as directed straight lines in the three-space through the origin of coordinates. However, if they were represented simply by a series of two-space plots (the usual mode of representation), the presentation would be misleading. What we would see would be a projection onto a plane (or two-space) of a vector located with reference to the coordinates of a three-space. Unless a vector happened to be orthogonal to the omitted dimension (i.e., located strictly in the plane), the resultant projection could easily be deceptive. This point is clear if one imagines a vector which in three-space extends almost directly out from the plane of the paper, and then imagines the projection of this vector on the plane or two-space. On the plane, the vector would appear to be very short or nearly invisible, but if one drew the conclusion that it had nothing to do with defining the space, one would be misled. To avoid this problem, we have reproduced as Fig. 5.1 the perspective drawing of the three-space configuration which appeared as Fig. 4.1, with the fitted vectors represented as lines with arrowheads in the positive

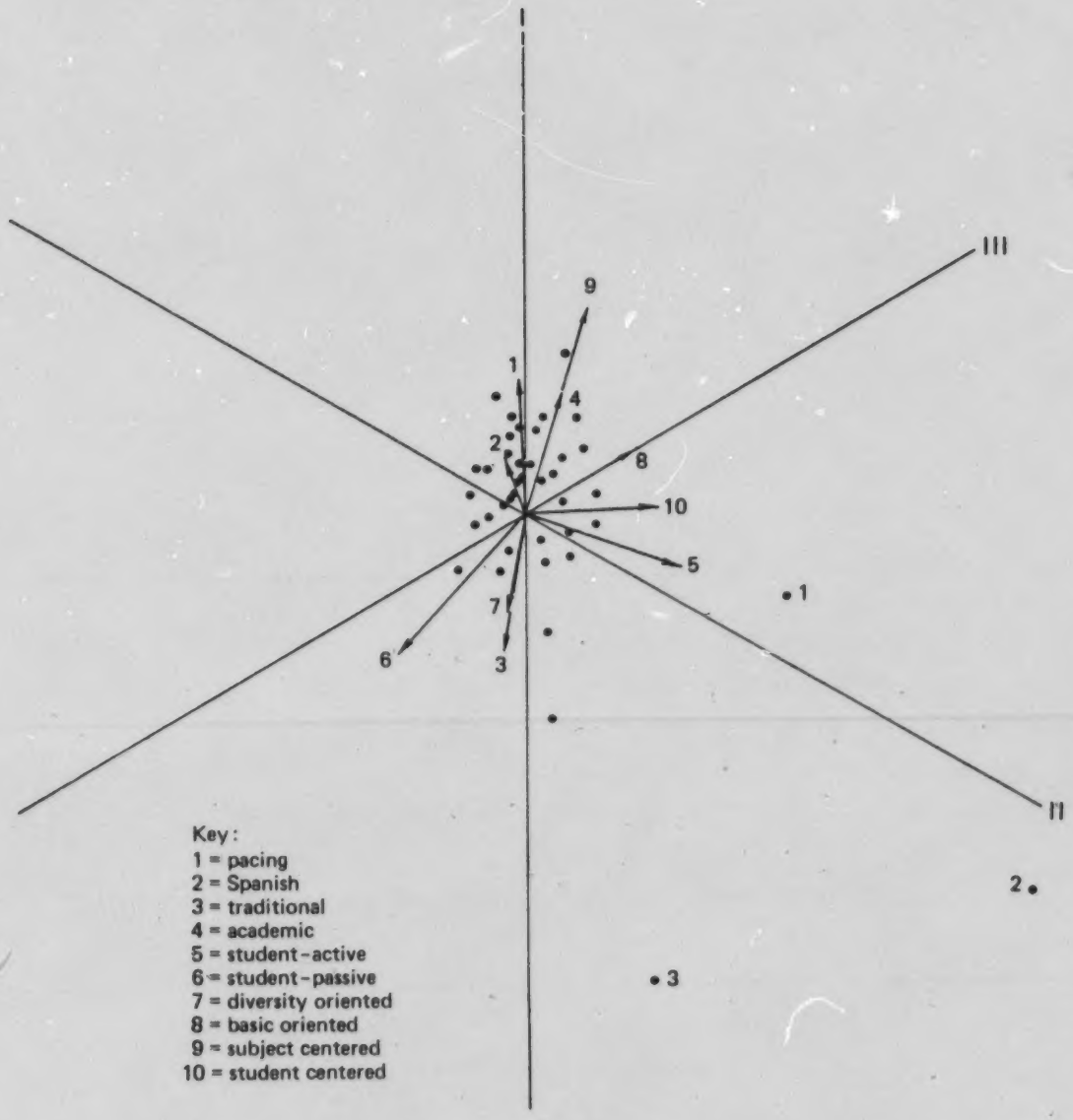

Fig. 5.1 — V classrooms in the 3-space configuration with interpretive vectors

direction. As before, points represent classrooms comprising the sample. The ten vectors are numbered in the same order as they are presented above, with a key providing variable names. Table 5.1 gives the correlation between the scale values and their projections.

Table 5.1

CORRELATION OF SCALE VALUES AND PROJECTIONS

PACING (1)	.62
SPANISH (2)	.36
TRADITIONAL (3)	.65
ACADEMIC (4)	.37
STUDENT-ACTIVE (5)	.44
STUDENT-PASSIVE (6)	.29
DIVERSITY ORIENTED (7)	.49
BASIC ORIENTED (8)	.31
SUBJECT CENTERED (9)	.43
STUDENT CENTERED (10)	.35

It should be apparent that, other things being equal, the higher the correlation the better the explanatory power of the variable, assuming that the direction of the relationship also makes sense. For example, if the most "diverse" classroom in our configuration had a low value on a scale purporting to describe "diversity orientation," we would have a serious problem of interpretation. On the other hand, if the bulk of the classrooms are relatively as homogeneous on these interpretive variables as they are on the DIVERSITY variables, this lack of variance will tend to lower the observed correlations. In that case, relatively high correlations are likely to result from the same outlier phenomenon we have observed in the configuration.

We begin with the PACING variable, which has the second highest correlation of the ten variables considered (.62). This is one of the rare variables whose vector does lie nearly in one of the planes.

Its angle with the second dimension is about 88°, making it nearly orthogonal to that dimension. Our expectation is that the more diverse classrooms, in our sense, should be highly self-paced. In classrooms characterized by a great deal of grouping, diverse activities and the like, we would hypothesize that several of the students or groups must be working more or less independently of continuous outside direction, both because this seems "theoretically" appropriate and because there would not be enough adults in the classroom to go around.

It is clear from Fig. 5.1 that this expectation is generally confirmed. Unit Three, which essentially defines the first dimension, seems to have the most positive projection on this vector, and Unit Two, even though it essentially defines the second dimension, seems to have the second most positive projection. These impressions are in fact supported by the data. Note, however, the PACING vector does not define any single dimension. It has a substantial angle with both the first and third dimensions, so that it defines, if anything, a definite direction in a plane. On the other hand, the third outlier, Unit One, which we saw as characterizing primarily the third dimension, has a relatively low (negative) projection on this vector. That is, Unit One is more nearly other-paced. In fact, its scale value is precisely in the middle of the possible range, or 1.5 in a scale that ranges from 1.0 to 2.0. The scale values for both other outliers are 2.0. (But this is also true of one unit which is not one of the outliers. The correlation is, as we say, not perfect.) In general, then, the results of the exercise with this variable are encouraging. Diversity in space and activities seems to be fairly strongly associated with student and self-pacing of the tempo and sequence of work.

Next, we turn to use of Spanish. We have already discussed our reasons for considering this variable. Briefly, it is said to be a distinguishing characteristic of certain Alum Rock classrooms, and theoretically it can be considered a likely accompaniment of individualization and diversity. With a correlation of .36, however, it is clearly not as strongly predictive of order as are some other

variables. Furthermore, it is not, as was the previous variable, conveniently located in a plane. Although it is most closely related to the third dimension (the angle is about 28°), it is certainly not orthogonal to the others (the approximate angles are 70° and 71° with the first and second dimensions, respectively). Of the three outliers, Unit One has the most positive projection on the vector (it is the highest of any point), while the other two are relatively low. However, of these three, only Unit One shows nonzero use of Spanish. Incidentally, Unit One ranks as one of the lowest of the ten classrooms that have nonzero values on the SPANISH variable--hence the relatively low correlation. This circumstance also makes reasonable the close identification of this vector with the third dimension, which is most clearly defined by Unit One.

Clearly, SPANISH is not empirically closely related to diversity as we conceive it. On the other hand, it is related to the Alum Rock nominal typology, although not quite as one might expect. Of the ten classrooms in the sample with nonzero Spanish usage, five are classified as BILINGUAL/BICULTURAL, and five (including the one with most recorded Spanish usage) as INDIVIDUALIZED, but not all BILINGUAL/BICULTURAL classrooms show a nonzero Spanish usage. Clearly, except for the last finding, this makes sense. Indeed, we have suggested that use of Spanish in a district like Alum Rock is congruent with individualization. On the other hand, none of these classifications is exclusive: Neither classification as BILINGUAL/BICULTURAL nor as INDIVIDUALIZED necessarily implies Spanish usage. Of course, we did not observe these classrooms exhaustively, and a finding of zero instruction in Spanish does not mean that it never happened. Nevertheless, the classrooms were in most cases observed for two entire days, which suggests that in those classrooms where Spanish was never heard, its use must be at least a rare event. Finally, as we have said, diversity from our viewpoint does not necessarily imply Spanish usage either theoretically or empirically.

Expectations about the third fitted variable, TRADITIONAL ($r = .65$), as shaped by theory, would be that the outlying units should be most negative on this scale. For Units Two and Three, this

expectation is confirmed; their projections are easily the most negative of all, and their scale scores are also the lowest. This is not true of Unit One; its projection is almost at the origin, and it has one of the highest scores on the scale. Such an outcome is very like the pattern obtained for PACING and begins to suggest that, if we thought in cluster terms, Unit One would belong in one cluster and the other two outliers in another. Recall, however, from Table 4.1, that Unit One is the least outlying among the outlier units, and in that sense is more like the bulk of classrooms than is either of the other outliers. On the present scale, almost all V classrooms including Unit One are quite high, with Two and Three forming a nonoverlapping distribution. This in part explains the relatively high correlation for TRADITIONAL, and begins to make sense of the configuration. Relative to the other V classrooms, Units Two and Three, which excel in diversity, are also entirely self-paced and just the opposite of traditional and nonindividualistic. On the other hand, Unit One, also somewhat differentiated, does show Spanish usage--unlike the other two--but otherwise is not so different from the remaining classrooms on the other two variables.

The variable called ACADEMIC ($r = .37$), if the fit were quite good, would be expected to show reasonably low scores for Units Two and Three with perhaps a middle to high one for Unit One. In fact, they are all in the middle to low range on this scale, with almost identical scores. In terms of projections, Units One and Two are outstandingly low with Unit Three near the middle, although all three are negative. This tends to explain the relatively low correlation for ACADEMIC. It also suggests that this scale is not measuring what we intend as well as does the other teacher activity scale, TRADITIONAL.

The next variable, STUDENT-ACTIVE ($r = .44$), is apparently somewhat better. Again we would expect the outliers to be relatively high, with Unit One perhaps slightly lower than the other two. This is just what we do find, with Unit Two receiving the highest score and exhibiting the second highest projection. Clearly, these two points account for the size of the correlation for STUDENT-ACTIVE, insofar as it is at all large. We expect Units Two and Three to be student-focused

relative to the majority of V classrooms and to Unit One, and that is what we find. Thus far, we seem to be getting a clearer interpretation of the first two than of the last one. For STUDENT-PASSIVE (r = .29), our theoretical expectations would reverse those for the previous scale. However, the lack of fit suggests that not much is explained by this variable. In fact, the scores for all outlying units on STUDENT-PASSIVE are in the low range and nearly identical; on the other hand, the projections follow the expected pattern exactly. But, given the lack of fit, this outcome is suggestive rather than definitive.

The DIVERSITY-ORIENTED scale (r = .49) shows interpretive promise, not simply because of its relatively high correlation but because it is nearly identical with the first dimension (its angles with the three dimensions in order are 9°, 87°, and 81°). We would theoretically expect, then, that Units Two and Three would be the highest on this scale, with Unit One high but lower than the other two. Again, this is exactly what we find, both for scale scores and projections. As App. D shows, the items of this scale are all aimed at assessing various kinds of diversity and individualization of student activities. The finding that this scale essentially defines the first dimension, as we saw Unit Three does, strongly suggests that what we conceive as diversity (i.e., what the DIVERSITY variables define) tends to be congruent with observers' summary judgments of diversity and individualization. It also suggests that, in some sense, Unit Three should epitomize both diversity and individualization in our sample. This suggestion is corroborated by notes from the observers' logs. The classroom is operated entirely by individual contracts between the teacher and each student. Moreover, Unit Three operates in this way for most classroom work. The exception is mathematics, which is more traditionally taught. For the BASIC ORIENTED scale (r = .31), our expectations are roughly the reverse of those for the previous scale although, because of the lower correlation, they are less strong. (It is fair to note that, as App. D makes clear, this is neither a strong nor a clear scale.) In fact, these expectations are not clearly confirmed, especially since Unit Two has the highest score of all 40

classrooms. Unit Three is among the lowest, with Unit One between. On the whole, it does not seem reasonable to put much weight on the BASIC ORIENTED summary scale.

The final two scales are interesting because they alone are not drawn from observational data, but rather are based entirely on teacher self-report data. Among other things, this implies that any congruence with results obtained from the DIVERSITY variables cannot be attributed to anything like halo effects, as some of the observational scales might be. In this way they are inherently more interesting than the preceding scales. The first of these, SUBJECT CENTERED ($r = .43$), as App. D shows, is marginally stronger than the STUDENT CENTERED scale. Moreover, its correlation is among the higher ones given. Our theoretical expectations are that Units Two and Three will be the lowest on the SUBJECT CENTERED scale, with Unit One higher but not the highest. In terms of both scores and projections, this expectation is precisely confirmed--a striking finding, given the disparate sources from which the data were drawn. The teachers of these classrooms, then, describe their classrooms in ways which suggest what the DIVERSITY variables are intended to measure: departures in several respects from the traditional classroom described earlier. Finally, the STUDENT CENTERED scale ($r = .35$) does not fit the space quite as well as its predecessor. Consequently, we believe that our theoretical expectations of highest scores for Units Two and Three and a lower one for Unit One will not be as soundly confirmed. They are not: The scores for all three outlying units are in the top of the range, and nearly identical, but a few V classrooms are higher. This distribution explains, in part, the relatively low correlation for the STUDENT CENTERED scale.

In summary, we have fit ten theoretically relevant variables to our configuration and they all accord approximately with our theoretical expectations. Of the ten, only one (DIVERSITY ORIENTED) is approximately linear with one of the configuration dimensions. As we saw, our expectations for that scale were fully corroborated by the data. In addition, the scales from self-report data, which provide another test of conceptions of classroom diversity, provided additional confirmation.

On the whole, we still do not know precisely what the configuration dimensions are, although the first (and strongest) is reasonably well specified in the concept of diversity advanced here. Moreover, it seems certain that diversity as measured by our DIVERSITY variables, which we took as necessary conditions for the occurrence of diversity, is in fact close to what both observers (cf. DIVERSITY ORIENTED) and teachers (cf. SUBJECT CENTERED) consider differentiation and individualization to be. Moreover, for what we considered the strongest interpretive scales, both conceptually and empirically, the expectations we held for the outlying units in a diversity space were generally confirmed. The patterns observed for Units Two and Three, the strong outliers, are very similar, suggesting that these two units in fact epitomize diversity in the V sample. The pattern obtained for Unit One is comparable but less similar. It is unfortunate that none of the interpretive variables is definitively congruent with the second or third dimensions, which suggests that future research might consider a wider selection of variables.

NONVOUCHER CLASSROOMS

As mentioned earlier, the fairly extensive investigation of comparisons among voucher and nonvoucher classrooms that we had originally intended was found to be unnecessary. Documentation for this assertion appears in Figs. 5.2, 5.3, and 5.4. These figures show the three two-space projections of a three-dimensional configuration in which the combined samples V and NV (total N = 74) were scaled.

In these figures triangles represent nonvoucher classrooms (the NV sample) while circles indicate voucher classrooms in V, those represented in earlier figures. It is clear immediately that, except for the three outlying classrooms previously identified, very little is different. In particular, like the majority of the voucher classrooms, the nonvoucher classrooms are clustered around the origin of coordinates and are inextricably intermingled with the nonoutliers. In short, the striking aspect of the nonvoucher classrooms is their similarity in spread and orientation--in lack of diversity as we understand it--to the body of voucher classrooms.

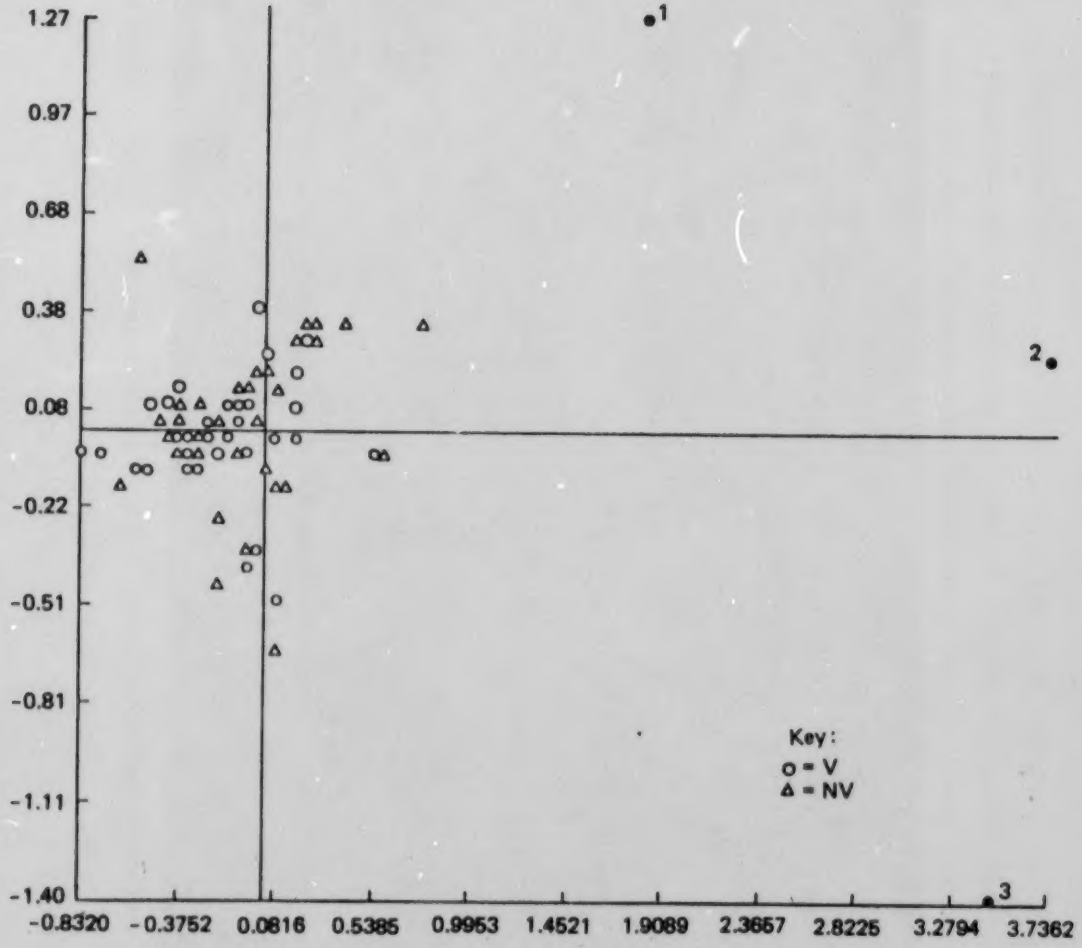

Fig. 5.2 — V and NV classrooms scaled in same space : I x II

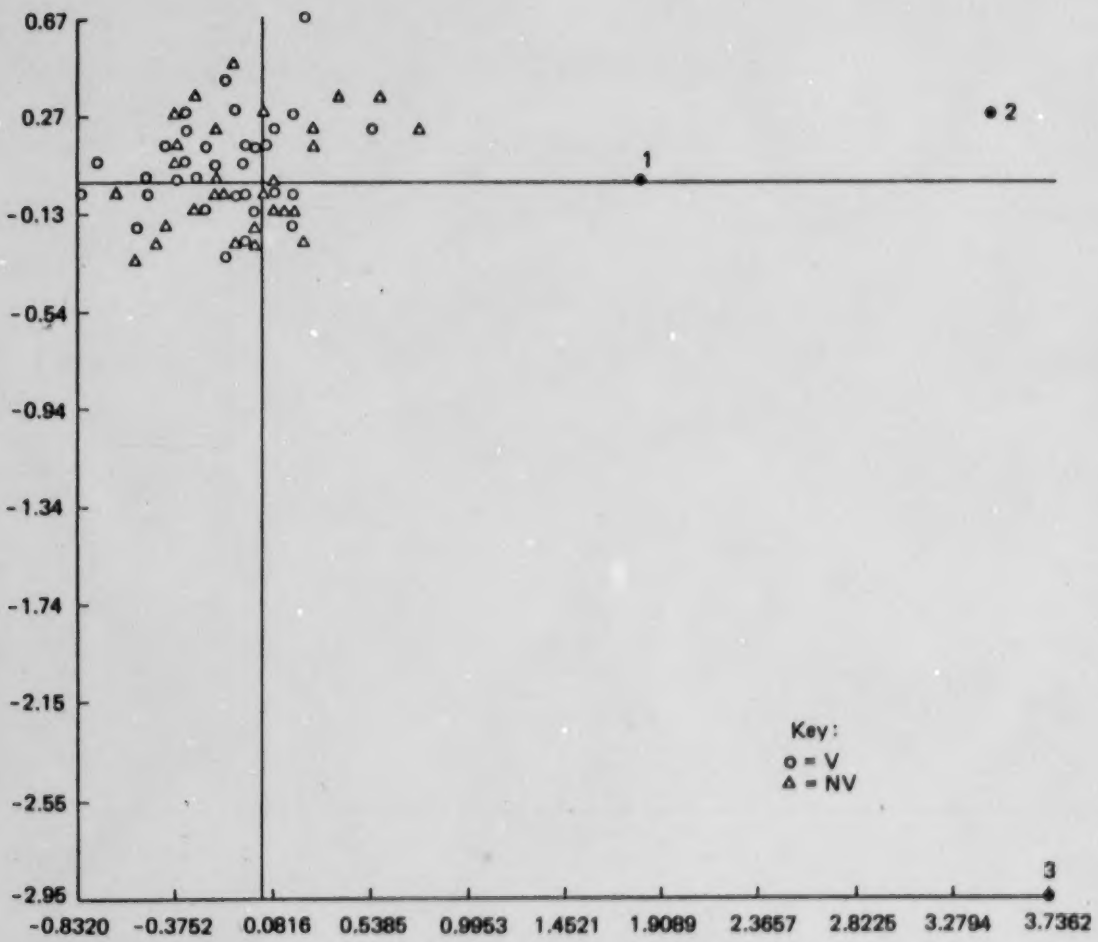

Fig. 5.3 — V and NV classrooms scaled in same space: I x III

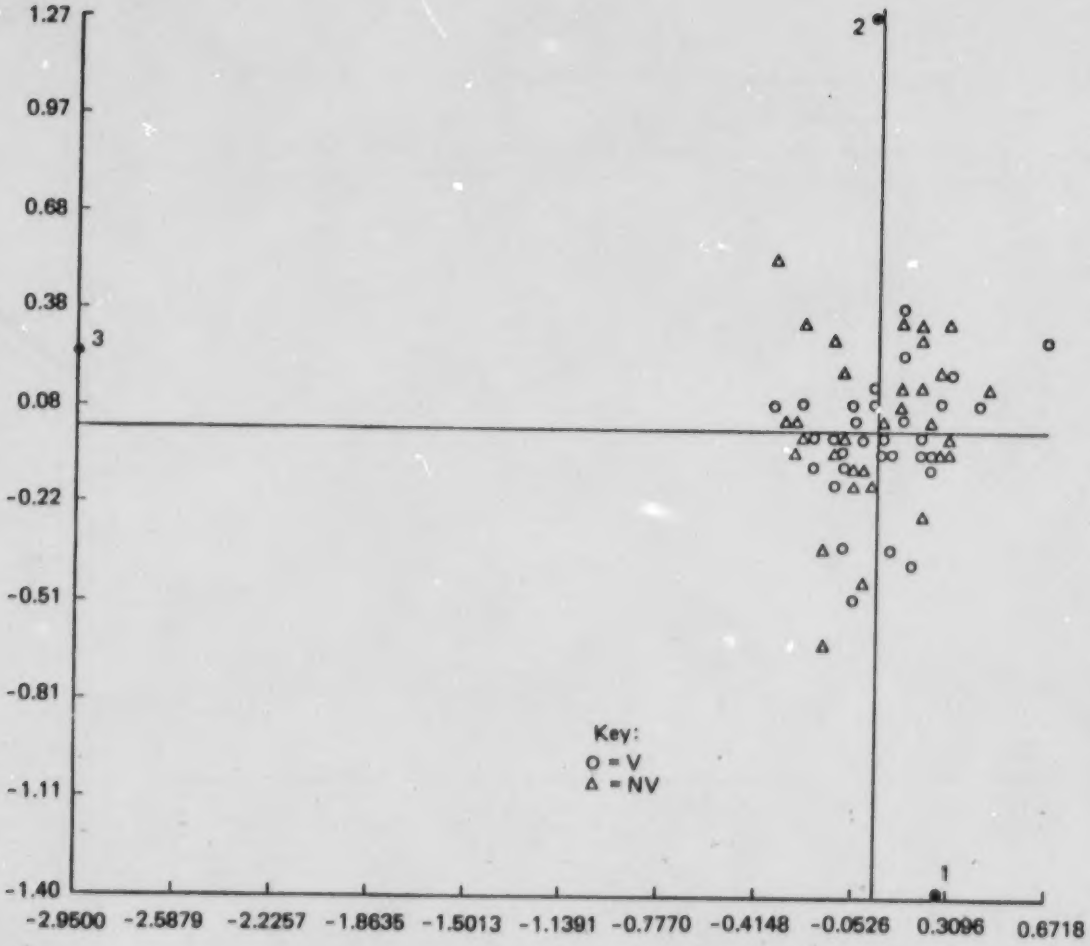

Fig. 5.4 — V and NV classrooms scaled in same space: II x III

On the other hand, the distribution of the outliers relative to the 74-unit configuration is not identical, except insofar as they remain outliers. Instead of being confined to a single dimension each as before, they now appear to vary on two dimensions each. (STRESS = .026 for this solution, negligibly different from STRESS in the V solution.) It is not clear what we are to make of this fact. Fitting the same ten explanatory variables into this configuration does not show any striking differences in correlations. Further, it does not seem worthwhile to make a detailed examination of the fit. The principal point--that of no essential difference between voucher and nonvoucher configurations--has been made.

SUMMARY OF ANALYSIS BY DATA SOURCE

With the interpretation of the obtained classroom space now complete, it would be well to review the many sources of information on which it was based.

Data resulting from the second round of observations were excluded from all major analyses, because of their low quality relative to data from the first round. These data quality problems, primarily involving a massive increase in the incidence of missing data and wild scores, resulted from an administrative budgetary decision to omit field review of observation instruments for omissions and errors during the second round of observation. Quality difficulties in the data did not become apparent until analysis was begun, at which point in time it was not feasible to attempt to get corrections for the raw data.

Considerable staff time and computing effort were expended to determine whether second-round observation data could be salvaged; our decision was that too large a percentage of scores would have to be estimated to render the resulting dataset trustworthy. However, examining correlations between valid second-round scores and first-round scores (in every instance in the .70s or better) led us to believe that--although pooling observations over two observation days would generate more stable data--analysis of first-round data alone would not lead to erroneous conclusions.

Classroom Observation Sheets

The Classroom Observation Sheets completed for the first round of observation provided most of the data for analysis:

The Classroom Roster was used to determine how well the class schedule was held, to select times for observation within an observation day, and to provide a selection switch, had separate analyses for differing subject matter been necessary. In addition, it supplied information about numbers of students and adults in the classroom, grade levels, and seating practices.

The Physical Environment inventory data served primarily to distinguish some V classrooms from other V and NV classrooms, and were a reflection of budget allocation decisions. As indicators of diversity, however, they were clearly inferior to data from the five-minute observation sequences.

The Five-Minute Observations generated the data of greatest value for scaling and interpreting the classroom space. The first five sections of the five-minute observation were used as a basis for constructing the DIVERSITY variables; the rationale and method for this procedure is given in Chap. 4. The BEHAVIORAL variables, constituting the remainder of the five-minute observation, were seen, relative to the DIVERSITY variables, as outcome or interpretive variables. These choices could have been reversed; that is, the multidimensional classroom space could have been constructed with BEHAVIORAL variables and interpreted via DIVERSITY variables. The two approaches in fact produce highly correlated constructions. But given the desirability of constructing the space on one set of variables and fitting interpretive vectors from information provided by the other, it seemed most appropriate to construct the space from the low-inference spatial/structural set.

Of the nine BEHAVIORAL variables, two were used in interpreting the configuration: SPANISH because of local interest in this variable, and PACING because of its immediate validity. For six of the seven remaining, some difficult dimensionalizing problems made them not immediately available for analysis; and careful exploration and resolution of these problems was precluded by scarcity of resources and

time pressure. However, an overall analysis based on a rough dimensionalization suggested both the validity of the major configuration and the susceptibility of the scaling problems with the BEHAVIORAL variables to solution.

Judgments supplied by the Classroom Log and Summary Scales were obtained explicitly to provide a basis, primarily using principal components analysis as a means of maximizing intrascale consistency and interscale distinctions after rational selection of scales, for the interpretation and explanation of configurations derived from the DIVERSITY variables MDS analyses. The resultant scales are described in App. D; some of them were not chosen for use because of their apparent lack of consistency, lack of relevance for interpretation, or both.

Teacher Questionnaire

The Teacher Questionnaire included considerable information to be used in future research for comparisons with data gathered from other school districts that offer educational alternatives. For the diversity study, data included in major analyses were the descriptive items on p. 1 of the questionnaire, and the attitude items (numbered 30 to 51). They, too, were intended to be used in the construction of interpretive scales, as described above. However, they are of special interest in this regard because the responses were generated by observed teachers rather than by observers. To the extent that researchers and practitioners provide consistent views of the learning environment, and to the extent that a model of the classroom space generated over one set of data is interpretable using data from an entirely independent source, our confidence in the findings is increased.

In sum, the Classroom Observation Sheets and the Teacher Questionnaire provided a rich base of information from which variables most suitable for constructing and interpreting a classroom space were selected, refined, and analyzed. As we suggest in the concluding section of this report, however, a wealth of data remains that could fruitfully be explored in future research.

Chapter 6

DISCUSSION

We began by pointing out that what we meant by diversity was not all that diversity might mean. Our analyses suggest that voucher classrooms on the whole were remarkably similar on the measures of diversity we employed. In short, they were neither very diverse nor very different among themselves. This does not imply that they are all nearly the same; rather, it means only that they are all nearly the same (relative to the two or three that are strikingly different) from an important, but not the sole possible, viewpoint. For example, it is clear that not all classrooms use the same sets of textbooks. Each classroom has a different teacher, a different collection of students, a unique flavor, and so on. In short, we make no assertions about diversity beyond what the data will permit.

With these caveats, however, we may discuss the question of why so many of the classrooms were relatively nondiverse, or so similar to each other. Or, to turn it around, we may inquire why very few of the classrooms were strikingly different. We have seen already that the answer does not seem to lie in mini-school membership. Each of the outlying units belongs to a mini-school whose other sampled classroom(s) lie(s) comfortably within the nondifferentiable swarm. Nor does the explanation lie in the Alum Rock typology; the same structure applies. We might hypothesize that the diversity which we found is attributable to the voucher demonstration. After all, we found no nonvoucher classrooms that lay outside the vague boundaries of the majority of voucher classrooms. This may be taken as some evidence for that position. On the other hand, in the nonvoucher sample, exceptions were likely to have been missed by sampling procedures, as we argued previously. We do not know whether we missed the one or two highly diversified, individualized classrooms in the sampled grade range in nonvoucher schools. We do not know whether they exist, a consideration that attenuates the weight of this finding.

Is there additional evidence bearing on this question? It seems not. If we knew the history of the teachers of the outlying classrooms, we could determine whether or not their classroom style and management predated the voucher demonstration, and this would count as evidence. However, we do not know that history, and have no access to it. Hence, we cannot conclude, but we can speculate. If someone were to raise the question whether the diverse classrooms found in the V sample could have occurred without the voucher demonstration, that possibility would have to be affirmed. After all, the practice of individual teacher/student contracts is not unique to Alum Rock, nor are the other differentiating features. Clearly, they can arise without the voucher demonstration. In short, if the demonstration had been a sharp spur to diversity, we would expect to find more of it than we did. Differences between voucher demonstration schools and others seem to reside far more in administration than in classroom practice. Another possibility, of course, is that what we have seen amounts to some sort of error--that the apparent outliers are not real cases, but artifacts of method or measurement. We have dealt with this possibility above, and have rejected it. We mention it here only for the sake of completeness.

So we are left with the fact that two or three classrooms in our sample are strikingly different in organization and operation from the remainder. This suggests that the single other known difference is the teachers who conduct those classrooms. In short, we may be dealing with what psychologists call individual differences, and it is often the case that their existence is a good deal clearer than their explanation. Our best guess, then, is that we are observing the results of unexplained teacher preferences for certain modes of organization and instruction that are not shared, at least not observably, by most of their colleagues. It is possible that the preference is shared, but is for some reason short-circuited on the way to implementation. However, we have seen that expressions of teacher preference embodied in scales constructed from teacher self-reports are reasonably congruent with practice as embodied in our observational results. Moreover, if this preference existed but had been for some reason

frustrated, it would seem that the voucher demonstration was just the opportunity to act on it. New approaches were explicitly encouraged and, up to a point, provided with financial backing. However, to repeat, this is speculation. One obvious next step would be to present to teachers some elaborated form of the question, "Why do you do what you do?" That is, we would like to have stepped beyond self-description to self-explanation. Such a step is proscribed by confidentiality agreements which prevent the linking of research results with individually identifiable teachers.

DIVERSITY RECONSIDERED

At several points in this report, we have cautioned the reader that where we speak of diversity we mean "diversity as we measured it." At these same points, we have discussed why we think the concepts and measures employed here are sound, and we remain convinced that they are. We emphasize that we do not contend that we have somehow described "true" diversity. What we think we have is a set of variables that measure the existence of conditions that we believe to be necessary for individualization of instruction and/or cross-sectional differentiation of activity. This is not the only way in which classrooms can differ, but it may be the principal or most important structural way.

Further, we make no judgment (mainly because we have no evidence) about whether diversity is a good thing. Obviously, the question is a very broad one, and we do not attempt to answer even specific versions of it. Many of these questions are, under present circumstances, unanswerable. For example, we would not want to try to decide whether students in diversified classrooms learn more. Some versions of the question may be more approachable. For example, it would not be unreasonable to attempt to determine whether students in alternative kinds of classrooms are more content, more stimulated, more bored, and so on. In any event, a greater range of educational variation would be required before any such research could be undertaken. In summary, we set out to measure something which we call, for short, diversity, and to determine whether classrooms in a certain

school district seemed to be characterizable in those terms. We have found that they are so characterizable and that our measures detect limited variation among classrooms.

FUTURE RESEARCH

At various places in this report, we have suggested extensions of our research approach which we consider interesting per se, and possibly even valuable beyond themselves. These suggestions were not exhaustive, nor are those to follow. For immediate purposes, it might be well to indicate that had we known the answer to one of the primary research questions--Are classrooms within mini-schools more nearly alike than those between mini-schools?--we would not have reduced our sample to V. We now believe that we could have included all sampled classrooms (even those in nonreplicated sets) without risk. In view of our commments about sampling in this kind of context, a reanalysis with the full sample would be desirable. Even better would be a restudy without sampling at all.

Furthermore, given the conclusions we have reached so far, a much more thorough exploration of spaces based upon alternative data sets is clearly indicated. Since our aims were limited, we confined ourselves to what seemed appropriate secondary analyses of, for example, the BEHAVIORAL variables and much of the self-report data. Clearly, a good deal more analytic work might be done. For instance, given our broad hypotheses about the DIVERSITY variables, it would be very interesting to explore in more detail the relationships between DIVERSITY and BEHAVIORAL spaces. The same is true of self-report spaces. While the use of the self-report data for interpretive purposes was enlightening, further exploration of the relationships between spaces based upon these data and those based upon observational data is clearly desirable--particularly because hypotheses in this area are not nearly so clear as in the former case.

We will not rehearse our discussion of the value of direct similarity estimation. Our findings based upon distance estimation from DIVERSITY variables reinforce the imputed value of such an approach. Moreover, we hope it is clear that replication of this work on new

samples, both for the light that may be shed upon the generality of the method and for substantive reasons, is desirable. One goal for this research is the construction of a method for fast and accurate assessment of classroom diversity which will produce reliable data. Finally, even without firm judgments about whether diversity is a good thing, explorations of both its antecedents and means for its encouragement are obviously desirable. Moreover, some beginnings of the assessment of the relationship of diversity to outcomes of interest could and should be included.

Appendix A

INSTRUMENTS

This appendix contains reproductions of the two instruments used for collection of data in the study of diversity in Alum Rock. The Classroom Observation Sheet appears first, and has been truncated in the following way: Five-minute observation forms are given only for the first ten minutes of an observation period. The complete instrument has four twenty-minute segments, or 14 additional five-minute observations identical to those reproduced here. Following the observation instrument is a complete copy of the Teacher Questionnaire.

CLASSROOM OBSERVATION SHEET

FACE SHEET

Tear Line

1-14/

OBSERVATION # ☐ *15/*

OBSERVER NAME _____ ☐☐ *16-17/*

OBSERVER PAIR ☐☐☐☐ *18-21/*

DATE OF OBSERVATION ☐☐☐☐ MO. DAY *22-25/*

TIME IN ☐☐☐☐ HRS. MIN. *26-29/*

TIME OUT ☐☐☐☐ HRS. MIN. *30-33/*

ACTIVITIES CODE

READING..... 01
ARITHMETIC..... 02
LANGUAGE ARTS..... 03
FOREIGN LANGUAGE..... 04
SCIENCE..... 05
SOCIAL STUDIES..... 06
STORY..... 07
SHARING..... 08
ART..... 09
MUSIC..... 10
PHYSICAL EDUCATION..... 11
RESTING..... 12
OTHER..... 13

-1-

CLASSROOM ROSTER

CARD 01 79-80/

SCHEDULE: ACTIVITIES PLANNED

		CODE	Hour	Min.	Hour	Min.	
1.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	21-30/
2.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	31-40/
3.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	41-50/
4.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	51-60/
5.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	61-70/ CARD 02 79-80/
6.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	21-30/
7.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	31-40/
8.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	41-50/
9.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	51-60/
10.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	61-70/ CARD 03
11.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	21-30/ 79-80/
12.	_____	☐☐	☐☐ :	☐☐	to ☐☐ :	☐☐	31-40/

STUDENT INFORMATION

1. Number enrolled ☐☐ 41-42/
2. Number present ☐☐ 43-44/
3. Students out of class for:
   - A. Staggered reading — A.M. ☐☐ 45-46/ P.M. ☐☐ 47-48/
   - B. Resource rooms, library, etc. ☐☐ 49-50/
   - C. Seeing a specialist ☐☐ 51-52/
   - D. Other, specify _____ ☐☐ 53-54/

   _____
4. Number in each grade:

Grade	☐ 55/	☐ 58/	☐ 61/	☐ 64/
Number	☐☐ 56-57/	☐☐ 59-60/	☐☐ 62-63/	☐☐ 65-66/

5. Graded..... 1 67/
   Ungraded..... 2
6. Preassigned Groups
   Yes..... 1 68/
   Sometimes..... 2
   No..... 3
7. Preassigned Seats
   Yes..... 1 69/
   Sometimes..... 2
   No..... 3

ADULT INFORMATION

Number of:

1. Regular Teachers ☐ 70/
2. Parent Volunteers ☐ 71/
3. Bilingual Adults ☐ 72/
4. Cross-grade Tutors ☐ 73/
5. Regular Aides ☐ 74/
6. Specialists or Resource Teachers ☐ 75/
7. Total Adults ☐ 76/
8. Other ☐ 77/
   Specify: _____

   _____

9. Total Duration of Class Day ☐ : ☐☐ 78-80/ CARD 04 79-80/
   HOURS MINUTES

PHYSICAL ENVIRONMENT

		PRESENT		USED	
1.	Clustered seating	☐	21/	☐	22/
2.	Room dividers	☐	23/	☐	24/
3.	Carpeting	☐	25/	☐	26/
4.	Arts and crafts materials	☐	27/	☐	28/
5.	Musical instruments	☐	29/	☐	30/
6.	Cultural displays	☐	31/	☐	32/
7.	Science equipment, objects	☐	33/	☐	34/
8.	Cameras, projectors, slide or film equipment, television, record player	☐	35/	☐	36/
9.	Individualized audio-visual equipment, tape player, materials	☐	37/	☐	38/
10.	Math or reading centers	☐	39/	☐	40/
11.	Other activity centers	☐	41/	☐	42/
12.	Math manipulatives, learning games	☐	43/	☐	44/
13.	Packaged learning kits	☐	45/	☐	46/
14.	Magazines, newspapers, other nontraditional reading matter	☐	47/	☐	48/
15.	Games, toys, indoor play equipment	☐	49/	☐	50/
16.	Posted classroom rules	☐	51/	☐	52/
17.	Posted schedules or assignments	☐	53/	☐	54/
18.	Student work displayed	☐	55/	☐	56/
19.	Other: specify:				
	_____	☐	57/	☐	58/
	_____	☐	59/	☐	60/
	_____	☐	61/	☐	62/
	_____	☐	63/	☐	64/

CARD 05 79-80/

-4-

17-20/

FIVE-MINUTE OBSERVATION: 1.1 15-16/ TIME: [ ][ ] : [ ][ ] HR. MIN.

1. Number of students not in groups: [ ][ ] 21-22/
   a. number tutored by an adult: [ ][ ] 23-24/
   b. number tutored by a peer: [ ][ ] 25-26/
   c. number with cross-grade tutor: [ ][ ] 27-28/
2. Number of student groups: [ ][ ] 29-30/
3. Number of students in the four most numerous groups: A. [ ][ ] 31-32/ B. [ ][ ] 33-34/ C. [ ][ ] 35-36/ D. [ ][ ] 37-38/
4. Group dissolves: [ ] 39/ [ ] 40/ [ ] 41/ [ ] 42/
5. DIVERSITY VARIABLES:

a. Location: (Circle all that apply)				
Regular desks or work area	1 43/	1 52/	1 61/	1 70/
Library area	2 44/	2 53/	2 62/	2 71/
Audiovisual center	3 45/	3 54/	3 63/	3 72/
Math, reading lab	4 46/	4 55/	4 64/	4 73/
Science center	5 47/	5 56/	5 65/	5 74/
Other special learning area	6 48/	6 57/	6 66/	6 75/
Floor	7 49/	7 58/	7 67/	7 76/
Other: _____	8 50/	8 59/	8 68/	8 77/
_____	9 51/	9 60/	9 69/	9 78/
b. Task: (Circle all that apply)	CARD 06	79-80/		
Reading	1 21/	1 31/	1 41/	1 51/
Mathematics	2 22/	2 32/	2 42/	2 52/
Language arts	3 23/	3 33/	3 43/	3 53/
Science	4 24/	4 34/	4 44/	4 54/
Social Studies	5 25/	5 35/	5 45/	5 55/
Other, academic	6 26/	6 36/	6 46/	6 56/
Music	7 27/	7 37/	7 47/	7 57/
Art	8 28/	8 38/	8 48/	8 58/
Cultural	9 29/	9 39/	9 49/	9 59/
Other, non-academic:	10 30/	10 40/	10 50/	10 60/
c. Materials: (Circle all that apply)				
Programmed learning materials, kits	1 61/	1 21/	1 33/	1 45/
Texts	2 62/	2 22/	2 34/	2 46/
Books other than texts	3 63/	3 23/	3 35/	3 47/
Films, slides, other visual media	4 64/	4 24/	4 36/	4 48/
Learning games, manipulatives	5 65/	5 25/	5 37/	5 49/
Workbooks	6 66/	6 26/	6 38/	6 50/
Locally developed materials	7 67/	7 27/	7 39/	7 51/
Arts and crafts materials	8 68/	8 28/	8 40/	8 52/
Commercial dittos	9 69/	9 29/	9 41/	9 53/
Paper	10 70/	10 30/	10 42/	10 54/
Other: _____	11 71/	11 31/	11 43/	11 55/
_____	12 72/	12 32/	12 44/	12 56/
	CARD 07	79-80/		

-5-

	Ⓐ		Ⓑ		Ⓒ		Ⓓ	
d. Mode: (Circle all that apply)								
Boardwork	1	57/	1	68/	1	16/	1	27/
Learn by doing	2	58/	2	69/	2	17/	2	28/
Discussion	3	59/	3	70/	3	18/	3	29/
Question/answer drill	4	60/	4	71/	4	19/	4	30/
Oral reading	5	61/	5	72/	5	20/	5	31/
Listening, watching	6	62/	6	73/	6	21/	6	32/
Machine-student work	7	63/	7	74/	7	22/	7	33/
Individual work	8	64/	8	75/	8	23/	8	34/
Waiting	9	65/	9	76/	9	24/	9	35/
Other: _____	10	66/	10	77/	10	25/	10	36/
_____	11	67/	11	78/	11	26/	11	37/
		CARD 08		79-80/				
6. BEHAVIORAL VARIABLES: (Circle one under each item for each group)								
a. Student-adult interaction	1	38/	1	47/	1	56/	1	65/
Student-student interaction	2		2		2		2	
Both	3		3		3		3	
Neither	0		0		0		0	
b. Student cooperation	1	39/	1	48/	1	57/	1	66/
Student competition	2		2		2		2	
Both	3		3		3		3	
Neither	0		0		0		0	
c. Students participating	1	40/	1	49/	1	58/	1	67/
Students disrupting	2		2		2		2	
Both	3		3		3		3	
Neither	0		0		0		0	
d. Students in leadership role	1	41/	1	50/	1	59/	1	68/
Adult in leadership role	2		2		2		2	
Both	3		3		3		3	
Neither	0		0		0		0	
e. Positive feedback given	1	42/	1	51/	1	60/	1	69/
Negative feedback given	2		2		2		2	
Both	3		3		3		3	
Neither	0		0		0		0	
f. Activity is individualized	1	43/	1	52/	1	61/	1	70/
Activity is group-structured	2		2		2		2	
Both	3		3		3		3	
Neither	0		0		0		0	
g. Adult with group or	1	44/	1	53/	1	62/	1	71/
Group alone	2		2		2		2	
h. Group is other-paced or	1	45/	1	54/	1	63/	1	72/
Group is self-paced	2		2		2		2	
i. Spanish is used or	1	46/	1	55/	1	64/	1	73/
Spanish is not used	2		2		2		2	

7. Number of outside interruptions: ☐☐ 74-75/ CARD 09 79-80/

-6-

17-20/

FIVE-MINUTE OBSERVATION: 1.2 15-16/ TIME: ☐☐ : ☐☐ (HR. : MIN.)

1. Number of students not in groups: ☐☐ 21-22/
   a. number tutored by an adult: ☐☐ 23-24/
   b. number tutored by a peer: ☐☐ 25-26/
   c. number with cross-grade tutor: ☐☐ 27-28/
2. Number of student groups: ☐☐ 29-30/
3. Number of students in the four most numerous groups: A. ☐☐ 31-32/ B. ☐☐ 33-34/ C. ☐☐ 35-36/ D. ☐☐ 37-38/
4. Group dissolves: ☐ 39/ ☐ 40/ ☐ 41/ ☐ 42/
5. DIVERSITY VARIABLES:

a. Location: (Circle all that apply)				
Regular desks or work area	1 43/	1 52/	1 61/	1 70/
Library area	2 44/	2 53/	2 62/	2 71/
Audiovisual center	3 45/	3 54/	3 63/	3 72/
Math, reading lab	4 46/	4 55/	4 64/	4 73/
Science center	5 47/	5 56/	5 65/	5 74/
Other special learning area	6 48/	6 57/	6 66/	6 75/
Floor	7 49/	7 58/	7 67/	7 76/
Other: _____	8 50/	8 59/	8 68/	8 77/
_____	9 51/	9 60/	9 69/	9 78/
b. Task: (Circle all that apply)	CARD 10	79-80/		
Reading	1 21/	1 31/	1 41/	1 51/
Mathematics	2 22/	2 32/	2 42/	2 52/
Language arts	3 23/	3 33/	3 43/	3 53/
Science	4 24/	4 34/	4 44/	4 54/
Social Studies	5 25/	5 35/	5 45/	5 55/
Other, academic	6 26/	6 36/	6 46/	6 56/
Music	7 27/	7 37/	7 47/	7 57/
Art	8 28/	8 38/	8 48/	8 58/
Cultural	9 29/	9 39/	9 49/	9 59/
Other, non-academic:	10 30/	10 40/	10 50/	10 60/
c. Materials: (Circle all that apply)				
Programmed learning materials, kits	1 61/	1 21/	1 33/	1 45/
Texts	2 62/	2 22/	2 34/	2 46/
Books other than texts	3 63/	3 23/	3 35/	3 47/
Films, slides, other visual media	4 64/	4 24/	4 36/	4 48/
Learning games, manipulatives	5 65/	5 25/	5 37/	5 49/
Workbooks	6 66/	6 26/	6 38/	6 50/
Locally developed materials	7 67/	7 27/	7 39/	7 51/
Arts and crafts materials	8 68/	8 28/	8 40/	8 52/
Commercial dittos	9 69/	9 29/	9 41/	9 53/
Paper	10 70/	10 30/	10 42/	10 54/
Other: _____	11 71/	11 31/	11 43/	11 55/
_____	12 72/	12 32/	12 44/	12 56/

CARD 11 79-80/

		(A)		(B)		(C)		(D)	
d.	Mode: (Circle all that apply)								
	Boardwork	1	57/	1	68/	1	16/	1	27/
	Learn by doing	2	58/	2	69/	2	17/	2	28/
	Discussion	3	59/	3	70/	3	18/	3	29/
	Question/answer drill	4	60/	4	71/	4	19/	4	30/
	Oral reading	5	61/	5	72/	5	20/	5	31/
	Listening, watching	6	62/	6	73/	6	21/	6	32/
	Machine-student work	7	63/	7	74/	7	22/	7	33/
	Individual work	8	64/	8	75/	8	23/	8	34/
	Waiting	9	65/	9	76/	9	24/	9	35/
	Other: _____	10	66/	10	77/	10	25/	10	36/
	_____	11	67/	11	78/	11	26/	11	37/
			CARD 68		79-80/				
6.	BEHAVIORAL VARIABLES: (Circle one under each item for each group)								
a.	Student-adult interaction	1	38/	1	47/	1	56/	1	65/
	Student-student interaction	2		2		2		2	
	Both	3		3		3		3	
	Neither	0		0		0		0	
b.	Student cooperation	1	39/	1	48/	1	57/	1	66/
	Student competition	2		2		2		2	
	Both	3		3		3		3	
	Neither	0		0		0		0	
c.	Students participating	1	40/	1	49/	1	58/	1	67/
	Students disrupting	2		2		2		2	
	Both	3		3		3		3	
	Neither	0		0		0		0	
d.	Students in leadership role	1	41/	1	50/	1	59/	1	68/
	Adult in leadership role	2		2		2		2	
	Both	3		3		3		3	
	Neither	0		0		0		0	
e.	Positive feedback given	1	42/	1	51/	1	60/	1	69/
	Negative feedback given	2		2		2		2	
	Both	3		3		3		3	
	Neither	0		0		0		0	
f.	Activity is individualized	1	43/	1	52/	1	61/	1	70/
	Activity is group-structured	2		2		2		2	
	Both	3		3		3		3	
	Neither	0		0		0		0	
g.	Adult with group or	1	44/	1	53/	1	62/	1	71/
	Group alone	2		2		2		2	
h.	Group is other-paced or	1	45/	1	54/	1	63/	1	72/
	Group is self-paced	2		2		2		2	
i.	Spanish is used or	1	46/	1	55/	1	64/	1	73/
	Spanish is not used	2		2		2		2	

7. Number of outside interruptions: [ ][ ] 74-75/ CARD 69 79-80/

-36-

CLASSROOM LOG

TEACHING STYLE

	NEVER	INFREQUENTLY	MODERATELY	OFTEN	
1. The teacher's questions are open-ended:	0	1	2	3	21/
2. The teacher makes use of student ideas in class (restates them, elaborates them, uses them in examples, asks further questions, etc.):	0	1	2	3	22/
3. The teacher employs direct (traditional) instructional methods (lecturing, demonstrating, and other explicit instructional techniques) with groups of students:	0	1	2	3	23/
4. The teacher's communication focus is individual students:	0	1	2	3	24/
5. The teacher's communication focus is groups of students:	0	1	2	3	25/
6. The teacher spends time interacting with other adults:	0	1	2	3	26/
7. The teacher provides students with feedback about their academic work:	0	1	2	3	27/
8. There is discussion of feelings, behaviors, beliefs, and attitudes:	0	1	2	3	28/
9. The class is working quietly, individually:	0	1	2	3	29/
10. The class works by discussion or other group-interactive processes:	0	1	2	3	30/
11. The class gives attention to alternative cultural themes:	0	1	2	3	31/
12. The class gives attention to fine arts:	0	1	2	3	32/
13. The teacher spends his/her time interacting with students:	0	1	2	3	33/
14. Curriculum is divided into subjects which are taught separately (vs. integrated curriculum):	0	1	2	3	34/

	NEVER	INFREQUENTLY	MODERATELY	OFTEN	
TEACHING STYLE (Cont'd)					
15. Team teaching is employed by teachers and/or aides:	0	1	2	3	35/
16. The class spends time watching films, television, or hearing stories, records:	0	1	2	3	36/
17. The class spends time in "fun" activities (games, songs, dramatic plays, etc.):	0	1	2	3	37/
18. The teacher is engaged in administrative, grading, or other non-interactive tasks:	0	1	2	3	38/
19. Other classroom adults are engaged in administrative, grading, or other non-interactive tasks:	0	1	2	3	39/
MANAGEMENT STYLE					
1. During recitation, students are recognized on a volunteer basis:	0	1	2	3	40/
2. During recitation, students are recognized in a patterned order:	0	1	2	3	41/
3. During recitation, students are recognized at random:	0	1	2	3	42/
4. Written work is assigned:	0	1	2	3	43/
5. Students' work is corrected by the teacher or another adult:	0	1	2	3	44/
6. Students' work is corrected by other students:	0	1	2	3	45/
7. Students' work is self-corrected:	0	1	2	3	46/
8. Order is maintained in the classroom:	0	1	2	3	47/
9. There is a smooth, efficient transition from activity to activity:	0	1	2	3	48/

MANAGEMENT STYLE (Cont'd)

	NEVER	INFREQUENTLY	MODERATELY	OFTEN	
10. The teacher gives negative feedback:	0	1	2	3	49/

Methods (Circle 1 or 0)	Yes	No	
a. Sent someone out of the room:	1	0	50/
b. Separated someone from his group or moved his seat:	1	0	51/
c. Gave a bad grade or other symbol:	1	0	52/
d. Made disapproving comments:	1	0	53/
e. Assigned extra work:	1	0	54/
f. Other			55/

Specify: _____ ☐☐ 56-57/

	NEVER	INFREQUENTLY	MODERATELY	OFTEN	
11. The teacher gives positive feedback:	0	1	2	3	58/

Methods (Circle 1 or 0)	Yes	No	
a. Gave someone a special privilege (e.g., first in line, a special book to read, etc.):	1	0	59/
b. Gave a good grade or other symbol:	1	0	60/
c. Made approving remarks:	1	0	61/
d. Released from a task or gave free time:	1	0	62/
e. Gave tokens, tickets or points·	1	0	63/
f. Other			64/

Specify: _____ ☐☐ 65-66/

-39-

STUDENTS

	NEVER	INFREQUENTLY	MODERATELY	OFTEN	
1. Students initiate questions, exchanges, discussions:	0	1	2	3	67/
2. Students are in the respondent role (answering questions, speaking when called on, etc.):	0	1	2	3	68/
3. Students are involved in self-instruction:	0	1	2	3	69/
4. Students select their own activities:	0	1	2	3	70/
5. Students select their own group:	0	1	2	3	71/
6. Students are expected to remain quiet unless called on to speak:	0	1	2	3	72/
7. Students are expected to stay where they are unless given specific permission to move about:	0	1	2	3	73/
8. Students stay in one place:	0	1	2	3	74/
9. Students spend a good deal of time waiting (for work to be okayed, for transition to the next activity, etc.):	0	1	2	3	75/
10. Students seek help from adults in the class:	0	1	2	3	76/
11. The class is task-oriented:	0	1	2	3	77/
12. Students engage in self-criticism or self-evaluation:	0	1	2	3	78/

UNUSUAL EVENTS RECORD:

CARD 70

79-80/

SUMMARY SCALES

1. There is an apparent in-class emphasis on basic skill learning.

   1 / 2 / 3 / 4 / 5 *21/*

2. There is an apparent in-class emphasis on self-teaching by discovery, exploration, doing.

   1 / 2 / 3 / 4 / 5 *22/*

3. There is an apparent in-class emphasis on self-other awareness.

   1 / 2 / 3 / 4 / 5 *23/*

4. There is an apparent in-class emphasis on cultural diversity.

   1 / 2 / 3 / 4 / 5 *24/*

5. There is an apparent in-class emphasis on fine arts.

   1 / 2 / 3 / 4 / 5 *25/*

6. The teacher seems to expect a lot of work to get done in class.

   1 / 2 / 3 / 4 / 5 *26/*

7. Students engage in a diversity of activities during the day.

   1 / 2 / 3 / 4 / 5 *27/*

8. Students work with a diversity of instructional materials during the day.

   1 / 2 / 3 / 4 / 5 *28/*

9. Student learning activities are individualized.

   1 / 2 / 3 / 4 / 5 *29/*

10. Students maintain interest in the class.

    1 / 2 / 3 / 4 / 5 *30*

CARD 71 *79-80/*

Appendix B

OPERATIONAL DEFINITIONS OF TERMS IN THE OBSERVATION INSTRUMENT

The text below includes excerpts from Rand's "Training Manual for the Alum Rock Classroom Observation Study," by P. Barker, T. K. Bikson, and J. Kimbrough (unpublished). These excerpts provide operational definitions for all of the important terms in the classroom observation instrument.

* * * * *

The most difficult and yet central term in the Five-Minute Observation instrument is GROUP. We will characterize a GROUP as precisely as possible, although by doing so we will not be able to eliminate all vague or borderline cases. So, while keeping the definition in mind, you will still have to rely at times on your commonsense judgment of whether or not a teacher is treating a collection of students as a GROUP. Most simply, a GROUP is a collection of two or more students (and possibly all the students), for whom the teacher has made a joint decision about what to do. The teacher is treating them as a single body in giving instructions about activities, rather than treating each as a separate individual about whom singular decisions regarding activities are being made. The first and most important cue to grouping is common location--the students are arranged in physical proximity. A second cue to grouping is common task--the teacher has given a joint set of instructions to students about what to work on. (However, it is not necessary that all students in the same group be working on the same thing. Sometimes they will be, and sometimes they may be working on different tasks.) A third cue may be found in the teacher's language. Sometimes the teacher refers to a collection of students by a group name (e.g., GROUP A), or sends a collection of students to a group work place (e.g., STATION ONE). On other occasions you can infer that a collection of students forms a GROUP because the teacher addresses them in second-person plural or otherwise indicates that she is

talking to them as a body. A fourth cue is interaction--students form a group if they are working as an interactive team (e.g., painting a group mural or doing an experiment). However, especially in traditionally structured classrooms, groups may not be interactive. Rather, each student in the group may work quietly and independently at his/her desk on math problems, a reading assignment, or what have you. (If all the students in the class are working in this fashion, they are scored as one group whose members = all the members of the class.) In summary, the key thing about a GROUP is that for decisionmaking purposes a collection of students is being treated as a single body. The teacher is making a decision about what to do for a number of students simultaneously. In the most easily recognizable case, a GROUP is given a common location and/or task. However, in other cases you will have to mentally review the above cues and then make a commonsense judgment.

A student is NOT IN A GROUP if decisions about what to do are being made for him/her individually, especially in regard to location and task. If a student is physically separated from the other students, and if an activity is specified explicitly for him/her, the student is NOT IN A GROUP. Further, a student is NOT IN A GROUP if he/she is receiving tutoring (whether the tutor is an adult or another student), or giving tutoring. Every student in the class must be counted as either a group member or as NOT IN A GROUP.

Given these notions of grouping, it is clear how the first two items in the Five-Minute Observation should be completed. That is, you will first count the students NOT IN GROUPS (in the sense just defined), and enter the number in the boxes to the right of item one. (In every case where there is more than one box, but you have only a one-digit number to record, use the right-hand box.) Among the students NOT IN GROUPS, count and record the following in a through c:

a. The number being tutored by an adult includes those receiving assistance on a 1:1 basis from either the teacher or another classroom adult.

b. The number being tutored by a peer includes those receiving such 1:1 assistance from another student who is a member of the class.

c. A cross-grade tutor is an older student who is not a member of the class and who is in the tutor role.

Having obtained information about students NOT IN GROUPS, you will attend for the remainder of the five minutes to students who are in groups. Item 2 indicates, at this moment, how many distinguishable student groups there are in the classroom. Having made this determination, you are now in a position to record the activity of each group. For this purpose, provide each group with an alphabetic designation (A, B, C, D). If there are fewer than four groups, you will not need all of the letters. If there are more than four groups, select the four largest groups to observe as A, B, C, and D.

Begin your observation with GROUP A. Fill in item three, the number of students in the GROUP as you start to observe it. Then proceed through all the diversity variables (5a-5d), and all the behavioral variables (6a-6i), recording information about GROUP A. When you are finished, count and record the number of students in GROUP B, then follow the same procedures for GROUP B. Continue in this manner until you have no more groups to record. Sometimes it happens that while you are in the middle of observing a group, the group dissolves (either the teacher disperses the group, the members finish their task and go on to do something else, or the period ends). In such cases, after you have recorded as much information as you could get, place a "1" in the "Group dissolves" box for that group. If a group has dissolved before you began to observe it, then the only information you can enter for it will be the "Group dissolves" item, item four. When item four is not marked, it indicates that you were able to complete your observation of the designated group, or else that there was never such a group at all. A score in the "dissolve" box represents that there was such a designated group, and that the group disbanded either before or during your observation of it.

Diversity Variables. There are four basic categories of diversity variables: location, task, materials, and mode. Under each category several items are listed and the procedure for recording them is always

the same. You will carefully observe the group under consideration, and then circle every descriptive item in the category which applies to that group. Many of these items are self-explanatory and need no definition. We will, however, provide observation guides whenever it seems useful.

Location. Every group must have a location, a physical context for its activity. Typically, a group's location is described by just one item, although it may happen that two adjacent locations are in use by a single group (e.g., a group using the science center and library area).

1. Regular desks or work area. This item refers to that portion of the classroom where most basic skill learning takes place, or where most seatwork is done. It is generally located toward the center of the room, with special areas located on the periphery. In a traditional classroom, this area would have desks (often, individually assigned desks for each student), and would be located near the blackboard. In nontraditional classrooms, nonassigned tables and chairs may be used instead. In any case, such an area must have work surfaces. It is conceivable that an entire classroom may be broken down into special learning areas, with nothing corresponding to regular desks or a multipurpose work area. In such a classroom this item would never be circled.
2. Library Area. The library area is set apart from the rest of the room and contains bookshelves or book tables. It may contain textbooks, although it is not just a text storage area. It must have nontext books as well (reference books, pleasure reading materials), and must have space for students to read or work. It may have tables, chairs, couches.
3. Audiovisual Center. Such an area is identifiable by audiovisual equipment set up for use. We do not mean an area where such equipment is stored, but rather one where either groups or individuals could work with the kind of equipment noted in the physical environment inventory as items eight and nine.

4. Math, Reading Lab. See physical environment item ten.
5. Science Center. This refers to a specific area set up for science projects with scientific equipment or objects and work space clearly evident (see also physical environment item seven).
6. Other Special Learning Center. This item refers to any other pre-established space set apart for academic or nonacademic activities, as included under the physical environment inventory item 11. It may be a multipurpose or single purpose area.
7. Floor. The floor should be scored as a group location only if it is not within any special learning area or center and is carrying out an activity. Most often, "floor" is scored if the teacher gathers students together and has them sit on the floor in an open area (alternatively, the teacher may ask them to bring their chairs to an open end of the room, in which case "floor" is also scored).
8. Other. Two spaces are left for you to write in group locations which do not properly fall under any of the descriptions above. Such locations will represent unique or different ways of using classroom space (e.g., a large cloak room or alcove might be pressed into service as an activity area, or a group of students might be located in a life-size model log cabin or teepee).

Task. Task descriptions represent what the students in a group are supposed to be doing when they are engaged in activity appropriate to the group. They may not represent what all of the students are, in fact, doing. Suppose, for instance, that you are observing a group of students to whom the teacher has handed out dittoed math problems to work. One of the students is, however, reading a comic hidden in his desk and two are racing to the window to see a fire truck speed by. You would, in this instance, score only "Mathematics." You would not also record "Reading" and "Other nonacademic activity." That is, you record only what counts as appropriate activity for that group of students. There are, of course, many occasions on which more than one task is accurately circled as appropriate. Suppose, for example, that the

teacher has asked students to finish whatever they had been working on that morning in the period left before the lunch break, or suppose the students are to do independent reading whenever they finish their math problems. In such instances, more than one task description should be circled. In general, definitions of task descriptions can be obtained directly from the discussion of the ACTIVITIES CODE above. Here we will clarify only a few terms which do not appear on that list.

Other, academic. Under this category include any discussion, lecture, or other assignment or activity that would be traditionally conceived as academic, and which is not represented in the preceding five items. Foreign language, English as a second language, economics, current events, and humanities are examples.

Cultural. This category is used to distinguish activities which are focused on cultures other than Anglo-American (cf. definition of cultural displays in the physical environment inventory). Such activities may be either academic (a subset of social studies), or nonacademic (perhaps music, crafts, cooking), but they are focused on understanding or appreciating other cultures.

Other, nonacademic. This category refers to tasks traditionally conceived as nonacademic (other than music or art). Circle this category for tasks such as resting, play, woodwork, cooking, and the like.

Materials. The materials variable represents learning media used by the students in pursuing their activities. Clearly, it is possible for one group of students to be using several materials in the list. On the other hand, it sometimes happens that students in a group are using no materials at all--as, for example, when they are simply watching and listening as the teacher reads a story, conducts a demonstration, or gives a lecture. If no materials are in use by a group, do not circle any items and do not write in the word "NONE." (This is the only diversity variable for which absence of response can be

correct.) Many of the materials items have already occurred in the physical environment inventory. In these cases we will refer you to the appropriate items. New entries will be characterized more fully.

1. Programmed learning materials, kits. Kits are defined as item 13 in the physical environment inventory. Programmed materials which are not parts of kits are also included, however.
2. Texts. A collection of books whose major function is to teach an academic subject. There are usually enough for everyone in the group, they are frequently hard-covered, and approved on a district-wide basis. This category does not include programmed texts for self-instruction.
3. Books other than texts. This category includes books and other reading matter not designed specifically for the purpose of reading instruction. First, all reference materials are included here. Second, books for pleasure reading are included. And third, all nontraditional reading materials also belong here (e.g., magazines, newspapers, and the like).
4. Films, slides, other visual media. This category includes the visual media described in items eight and nine (physical inventory).
5. Learning games, manipulatives. This category includes items defined in number 12, physical environment inventory.
6. Workbook. Workbooks are paperback books containing exercises for students to complete by writing in the books (although, in fact, teachers may ask them not to write in the workbooks).
7. Locally developed materials. This item refers to instructional materials which are obviously not commercially produced and which have been made by the teacher or aides (e.g., dittoed problems), by parents (e.g., a science project), or by students themselves (e.g., cut-out geometric shapes).
8. Arts and crafts materials. See physical environment inventory, item four, for standard art-craft materials. This category also includes other materials which are used in an art or

craft activity (e.g., newspapers for papier mache, magazines for collage material, wire for metal sculpture.

9. Commercial dittos. Single sheets of problems or exercises, much like teacher-made dittos, are included here; however, these sheets are clearly commercially produced.
10. Paper. This item refers to paper (lined or unlined, tablet or loose-leaf, either belonging to the student or distributed by the teacher), on which students write or cipher.
11. Other. Two spaces have been left for you to indicate use of instructional materials not represented in the above list. (Do not include pens or pencils.)

(Please note that sometimes learning kits include a programmed text, a workbook, dittos, activity cards, and the like. Do not separate kit items and score them individually. Enter only one description for the kit.)

Mode. The mode variable attempts to describe the manner in which students are working. The same activity can be carried out in any number of ways. For example, social studies could be undertaken by group discussion, by role play, by a question-answer drill, by lecture, and so on.

Keep in mind, however, the "mode" represents mode of activity of the students in the group you are observing. For example, "lecturing" is not listed as a mode because students virtually never do this. The teacher may teach by lecturing, however, in which case you would describe the mode of student activity as listening. At least one item under "mode" must be circled, and more than one may be circled.

1. Boardwork. This mode represents students' use of the blackboard for drawing, writing, or calculating. Often some of the students in a group use the board while others watch. In this case you would record both boardwork and watching. If the teacher or other adult is the only one using the board, do not score boardwork under mode because mode refers to student activity.

2. Learn by doing. This item refers to activities in which students learn by doing some sort of motor, perceptual, or symbolic performance other than traditional reading, writing, ciphering, and answering or discussing. For example, students may learn by role play, by using puzzles or games, by cooking, by making an imaginary investment or budget, and so on.
3. Discussion. Discussion occurs when students are encouraged to express their own opinions or attitudes, pose questions, give reasons or explanations, give criticisms or evaluations, and so on. In discussion, whether the subject matter is academic or nonacademic, open-ended questions are often involved, and a single "correct" answer is usually not sought.
4. Question/answer, drill. This mode of activity is fairly structured. Students are expected to give short, objective answers to questions, and responses are either correct or incorrect. An example might be naming the capitals of states, naming the original colonies, and so on. Under drill would be included saying multiplication tables, responding to flash cards, and so on.
5. Oral reading. Reading aloud, either singly or in unison, either to each other or to an adult, is included here. Any printed material may be involved (books, charts, alphabet cards, etc.).
6. Listening, watching. In this mode, students are passive/receptive. They are in the audience role, giving attention to an adult or another student, or to some other presentational medium (record player, film).
7. Machine-student work. Here is included any interaction between a student and a machine in the performance of an activity. Examples include any individualized audiovisual equipment, typewriters, calculators, and other pieces of educational technology. It does not include noninteractive machines (films, television), and does not include scientific equipment (microscopes, prisms).

8. Individual work. In this mode, students are working quietly and relatively independently. An example is the kind of "seat-work" often done in traditional classrooms. Activities in this mode are carried out at the regular desk or work area.
9. Waiting. This behavior mode occurs when students are waiting for assistance, for approval, to check something out, to be dismissed, and so on. During the wait, no other task is being done.
10. Other. Two spaces have been provided for you to record modes of activity which you observe and which do not fall under any of the categories provided.

Behavioral Variables. There are nine sets of items classified as behavioral variables, or descriptions which tap important aspects of student group behavior. As with the diversity variables, you must score each item for each group. For the behavioral variables, only one number can be circled in an item set. For most of the behavioral items you have four choices, consisting of two positive alternatives along with a "both" and a "neither" response. However, the last three items under this heading present just two alternatives, one of which must be chosen to characterize the observed group.

a. Student-adult or student-student interaction is characterized by active conjoint participation. The term "interaction" is the key. In adult-student interaction, the adult may be a teacher, aide, or other classroom adult, but the adult must be actively participating with the students in the group, either as a leader or in a peer-like role. Do not count testgiving, silent supervision, or disciplinary remarks as interaction. Student-student interaction is defined similarly. Do not score verbal or physical abuse among students as interaction. Both such interactive styles may occur in a single group (circle 3), or the group may be noninteractive (circle 0).
b. Student cooperation entails student-student interaction. In addition, students must be working together toward the completion

of a designated task; they are collaborating for a common goal or purpose, and no one finishes as a winner or loser. Student competition may or may not be interactive; however, the activity must involve a desired goal which is sought by students individually and cannot be attained by all of them collectively (e.g., seeing who can complete a task first, say the correct answer first, get the most correct, write the neatest, etc.). Score a group as competitive only if this orientation is clearly evident. It is doubtful that groups will be observed to be both competitive and cooperative; however, it is entirely possible that neither orientation characterizes a group.

c. Student participation entails that the students are "doing" rather than passively waiting, or being idle and restless. Participation includes performance of any behavior appropriate to a task (being engaged in the activity at hand, using the assigned materials, exhibiting the intended interaction mode). In contrast, disruptive behavior is behavior which requires (or would elicit) intervention from an adult. It includes any behavior inappropriate to and impermissible in the context of the activity (e.g., talking during a quiet period, jumping up and down during desk work, hitting or kicking, and so on). Please note that a behavior is disruptive only if it is against the implicit or explicit rules of the classroom. Thus, what seems disruptive to you is not to be scored as disruptive if it is a style of behavior clearly permitted by the teacher. Some groups may be characterized by both participatory and disruptive behavior. The "neither" category includes waiting, boredom, idleness, restless nondisruptive activity, having "nothing to do."

d. A group has a student in the leadership role if the student has been selected, elected, or has volunteered to be in charge of the activity. His/her role is in some sense more directive or responsible than the roles of the rest of the students in the group. More than one student may be involved (e.g., co-captains,

co-panelists leading a discussion, etc.). Do not score students giving a test to other students as students in the leadership role. That an adult is with the group does not necessarily mean that the adult is acting as leader. (Clearly, in the long run the adult has supervisory responsibility and may take over the leadership role if the group becomes chaotic. However, it is entirely possible for the adult to allow students to retain the leadership role for some portion of the classroom day.) The adult is acting as leader of the group when the adult takes active charge of the task. The "both" category can be employed only when both a student and an adult are leaders (either as co-leaders or sequentially). "Neither" describes a situation in which there is no leader--for example, when group members are working individually and when the adult present is simply monitoring, rather than leading, their activity.

e. This set of items refers to the responses teachers or other adults make to student performance, where those responses are critical, evaluative, or corrective. Feedback may be given to the group as a whole group, or it may be given to individual students. Moreover, it may pertain either to academic activities (e.g., thought problems that were correctly solved), or to personal-social behavior (e.g., the group worked the problems quietly and responsibly). If you observe that virtually all feedback is positive, circle 1. If it is virtually all negative, circle 2. If it is both, circle 3. If you do not perceive any feedback given, circle 0.

f. Individualization here inquires whether, even though students are grouped, their activities are mainly individually organized or mainly group-structured. If students in a group primarily work at their own pace, often on different tasks or else by using different materials or different parts of the same texts, and have different objectives, their activity would be regarded as individualized. If, in contrast, you observed the group to work interactively, or at a common pace on a common task or with common objectives, then the activity would be regarded as

structured for the group as a group. Again, it is possible that both descriptions equally characterize the group. However, it is difficult to imagine a case in which none of these three alternatives were appropriate.

g. This is the first strictly either/or item in the list of behavioral variables, and it refers simply to whether the group you observed worked alone or instead was accompanied by an adult (either the teacher, an aide, a parent volunteer, or cross-grade tutor). An adult is "with the group" if the adult circulates around the room, briefly stopping here and there to make comments, or if the adult merely calls out an instruction or warning to the group, but is not physically present with the group.

h. Pacing refers to the rate of progress through a task. Work is self-paced if the rate of progress is determined by the students (either individually or as a group). It is other-paced when either an adult or a machine paces the task. For example, if students begin a task at the place where they left off, working at their own speed, the task is self-paced. A task is self-paced if it is jointly managed by the group (as in producing a group mural). In contrast, a typical example of an adult-paced task is a reading group in which all students start at the same place, all students are paced through the reading by the teacher who calls on them and expects those not reading to follow along together. Or a task may be other-paced if the response rate is determined by a teaching machine.

i. The last item in this list taps use of Spanish. If you perceive that Spanish is appropriately used by students in the group, score a one. Here "appropriately" means that Spanish is permitted and intended as a language of instruction, and is being used in an academic activity rather than just colloquially and privately among students. Use of Spanish may be written or oral, and use of Spanish reading materials by students is included. If you do not perceive the use of Spanish in written or oral form by students, score two.

Finally, the last variable in the Five-Minute Observation instrument, question seven, asks for the number of outside interruptions which occurred during that observation period. By an "outside" interruption is meant any announcements made by intercom or by a person entering the classroom who is not ordinarily a student or adult in that classroom. Also counted for this purpose are any other entries into the classroom by outsiders (school staff, students), with messages for the teacher. Such events should simply be tallied as they occur, at the bottom of the page. At the end of the five-minute period, enter the number representing total outside interruptions in the boxes to the right of the item.

Classroom Log

The classroom log is a series of statements representing classroom procedures. These statements will be true, in varying degrees, of the classrooms you observe. To complete the log properly, you need first to become familiar with the kinds of classroom procedures represented in the log. For convenience, we have divided the statements into sets related to teaching style, classroom management, and students. While you do not have to memorize the items in each set, you do have to become well acquainted with them so that you will know what to attend to in the classroom. During times when you are not completing the five-minute observation instrument, you must pay attention to the classroom process and watch especially for occurrences of the kinds of events represented in the log. At the end of the morning's observation, you are to fill out the log.

Statements in the log require responses in the form of relative frequencies, ranging from a low of 0 to a high of 3. If the sort of event described in the statement never occurs (i.e., was not observed to occur at all), circle the 0. If the event described occurs relatively infrequently, score a 1. If the event sometimes occurs, if it happens in the described way as often as not (i.e., it happens moderately frequently), score a 2. Finally, if the event always, or often, occurs as described by the statement, score a 3. Obviously, relative judgments of this sort involve a comparison. You may initially feel that you do

not have enough experience to make such comparisons. After the training period, your experience and your confidence will have increased, although some subjectivity is always going to be involved. The best guide we can provide is to weigh the event against the class of background events of which it is part. For instance, consider statement 1: The teacher's questions are open-ended. In responding to this item, consider all the questions you have heard the teacher ask. If she either asks no questions, or none of them are open-ended, score 0. If, among all her questions, a very few are open-ended and the vast majority are objective, score 1. But if, in contrast, the teacher most often uses questions that are open-ended, then score 3. That is, you respond to the item by considering the teacher's question-asking as a whole, and recording a score that represents what proportion of the questions have the form described.

Statements addressed to teacher behavior refer literally to the teacher or teachers in the classroom. They are not intended to represent the behavior of other classroom adults unless the term "adult" is explicitly used. In general, the statements are self-explanatory. We will discuss below only a few statements from each section of the log.

Teaching style. Statements in this section are generally focused on the teacher in the instructional setting. Some of them describe teaching styles apparently contrary to one another (e.g., items 4 and 5), and we have tried to group these to facilitate your formulation of responses. Do not infer that because the statements are discrepant you cannot respond positively to both. For instance, the teacher could spend half of her communication time in 1:1 discussion with students and the other half of it with group-directed comments. In such a case, you might well score a 2 for both statements. On the other had, if virtually all the teacher's statements were group-focused and only a few individualized, you would score a 3 for item 5 and a 1 for item 4.

6. This item refers to teachers' interactions with other classroom adults in contexts that do not include students. That is, do not include here instances in which a teacher and aide, for example, interact with one another in the course of leading

a science project with students, or while working with a drama group, and the like. Include here only instances in which teachers and other adults are interacting with one another and not with students.

8. This item is intended to tap discussions of interpersonal events--how students feel about themselves, how they feel about one another, how they see each other, what they believe about other people, how they interpret the behavior of others toward themselves, and so on. In contrast to discussions of ideas and beliefs about academic subject matter, this item refers to classroom treatment of noncognitive, socioemotional themes.

11. Alternative cultures, in this item, include any cultures other than Anglo-American, as we indicated in the physical enviornment inventory. By "giving attention," however, we mean actively referring to such cultures--either by making an alternative culture the explicit subject of a period or by making references to alternative cultures in the context of other lessons. "Giving attention" does not include simply having cultural displays in the room.

12. "Giving attention" to fine arts similarly means active attention. It does not include simply having posters, books, or other representations of the fine arts in the room. Fine arts include painting, sculpture, literary works, and music and dance for appreciation and instruction; it also includes ordinary art-craft activity that excludes art appreciation, and it includes music activities for entertainment.

13. The notion of teacher-student interaction involved here is that described for the five-minute observation above (behavioral variable (a)). That is, it refers to active participation of the teacher with students. It does not include silent supervision, test giving, lecturing, or occasional disciplinary comments; it means co-participation.

17. In this item, "fun" activities may include games that students can learn from (e.g., Monopoly, ordinary card games); but they are not intended to include learning games, i.e., games manufactured specifically for educational purposes.

Management style. Statements in this section of the Classroom Log attempt to describe order, structure, and formal evaluation procedures. It is, of course, somewhat artificial to separate these from accounts of teaching style since the two are clearly related. However, statements in the management section seem to be generally more relevant to maintaining academic organization than to instructional practices per se.

1. Recitation is used here to include question/answer, drill, sharing, discussion--in sum, any sort of occasion during which students speak in class (not including occasions on which students speak privately to teachers). This item indicates that students are called on to speak if and only if they volunteer by raising their hands or otherwise indicating that they would like to speak.
2. Students are recognized in a patterned order for recitation if the teacher has a systematic method of calling on them--either alphabetically, by rows, by groups, by stations, or in some other predictable fashion.
3. Students are recognized at random if they are simply picked out and asked to respond; they have not volunteered, and the recognition order is not patterned in advance. They are just requested to respond.
4. In this item, "written work" is a phrase used to designate any kind of assignment (reading, mathematics, language arts, and so on) which the student is expected to complete. The assignment may be time-limited in class (as a test or other piece of work to be turned in at the end of a period), it may be due at the end of the week, or it may be homework. What is common in all these instances is that the student is expected to complete a specific product.
8. "Order" in the classroom refers here to the maintenance of structure or procedure. There may be written rules, verbally referenced rules, or simply implicit rules. In any case, there is little confusion, a low noise level, students appear to know what they should be doing, and classroom practices seem to be tightly ordered.

9. Transitions from activity to activity are smooth and efficient if they take little time, do not involve a lot of waiting or disruption, and it is evident that students and adults have a clear picture of what is going to happen next. Not much time and motion is wasted in the transition.

10. This item and the one following it differ from the others in the management section in this way. First, there is a general item inquiring about the relative frequency with which teacher feedback to students is negative or positive. The general statement refers to all feedback given, and inquires about how it is apportioned. After the general item there is a list of methods for delivering negative or positive feedback. If you have observed the teacher employing such a feedback method, circle 1 (for yes); if you have not observed the use of the method named, circle 0 (for no). Among the negative feedback methods, please note that item c includes not only grades but other devices such as a frowning face, a standardized bad-feedback note (e.g., "Susan was a yacky-bird today"), or other formal notice of poor performance. Method d, disapproving comments, is intended to include threats and warnings and grimaces, as well as other expressions of disfavor. The "other" category includes technique used by the teacher to convey negative feedback which we have not listed.

11. Complete the general item here as for 10 above; i.e., in relation to the total amount of feedback given by the teacher, what proportion of it is positive. Then indicate, as with item 10, the particular methods the teacher uses to convey positive feedback. Method "e" here is intended to capture the kinds of feedback employed in token economies and other behavior shaping approaches; adopting only positive reinforcers, these approaches dispense units of reward such as tickets, points, tokens, and so on (which are often accumulated and redeemed for a larger reward).

Students. The third section of the Classroom Log comprises statements about students. Again, we acknowledge that it is difficult to characterize student behavior independently of the style of classroom instruction and management, so the separation of sections is artificial. Consequently, we would like you to respond to the items here as they characterize the classroom you observed, independently of whether you think the characterization reflects student behavior or really says more about how the teacher has organized it.

1. The first statement here inquires about the relative frequency with which students are in the initiator role during the times when they speak.
2. The second item, in contrast, asks about the relative frequency with which students are in the respondent role during the times when they speak. They are in the respondent role if classroom adults are the ones who pose the questions or determine what are the issues to be discussed.
6. This item does not imply that the classroom is always quiet, and in fact says little about noise level. Rather, it is focused on whether it is a behavioral norm that students do not talk unless called on. In fact, students may violate the norm, but you should be able to recognize whether the classroom does have such a norm and whether students are expected to get permission before speaking.
7. This item, like 6 above, refers to classroom norms about freedom of movement. It does not ask how much student movement there is (this is tapped in item 8); rather, it attempts to find out who directs the movement. If students may always move about at will, score 0; if students must always have adult permission before getting up from their seats or going to another center, score 3. Intermediate cases are decided on the basis of the relative freedom available to students to determine their own movement in the classroom.
12. Students may engage in self-criticism or self-evaluation in respect to either their academic work or their social or personal behavior. Signs of such activity include a self-reflective,

analytical, or interpretive attitude toward what either they or other students have said, written, read, or done. It is also indicated when students provide feedback to themselves or to other students. It is, however, to be distinguished from boastings, self-deprecations, and ingratiating or hostile remarks to others.

Unusual events record. The Classroom Log concludes with a blank space allocated to the recording of unusual events. This space allows you to make note of anything that occurs during the observation day which either has not been scheduled, is atypical or unique, or might affect the recorded observations in an unusual way that we would find difficult to interpret if we did not have an account of what had happened. For instance, a fire drill might interrupt your observations. Or, someone in the class might suddenly become ill, a short-term chaos ensuing. Or a severe disciplinary problem might arise, causing a disruption of classroom activity. In short, any number of events could alter your observations. The unusual events record allows you to notify us informally of what has occurred, so that we can appropriately interpret the data.

Summary Scales

This final part of the classroom observation instrument is reserved for you to record your own informed perceptions. Filled out only once, at the end of the observation day, it represents the summary viewpoint of an experienced observer. The instrument parts preceding it are broken down into numerous detailed objective judgments, which we will attempt statistically to unify and interpret. But we need also to know what these details add up to in commonsense terms. We need to know what, in summary form, is evident to someone who enters the classroom and looks about for its distinctive features. For this reason, we have included 10 statements derived from the varied objectives that mini-schools have posed for themselves. We would like you to give the classroom you have observed a score from 1 to 5 on each of these scales. The number represents the extent to which the statement characterizes your summary perception of the classroom at the end of the observation day. A score of 5 means that the statement is very, or extremely, character-

istic of the classroom; a score of 1, in contrast, means the statement is very, or extremely, uncharacteristic of the classroom. The middle value, 3, is used to represent an intermediate judgment--the statement is not especially characteristic and not especially uncharacteristic of what you observed. Finally, a score of 4 indicates the statement is more characteristic than a neutral value of 3 would suggest, while a score of 2 indicates the statement is less characteristic than the neutral value of 3 implies. The guidelines below should be helpful in filling out the summary scales.

1. The concepts involved in "basic skill learning" refer at least to the academic subjects of language arts (reading and writing skills) and mathematics; there may be reference to other academic subjects as well, e.g., science, geography, history, etc. All mini-schools teach these basic skills; on the basis of their descriptions, they differ in how they teach these skills. If the class emphasizes basic skill learning, then the approach is (1) direct and (2) teacher-mediated. A paradigm will be useful here. (1) Consider a direct approach in teaching mathematics: One may teach via a textbook, packaged learning kits, math manipulatives, puzzles and the like; moreover, there may be a variety of instructional approaches including interactive and individualized approaches in addition to the traditional classroom scene. In contrast, an indirect approach would involve teaching mathematical concepts and operations while engaged in cooking, shopping, taking part in fund raising drives, etc. (2) By "teacher-mediated" we mean simply that the teacher is the directive focus such that there is a minimum of student-student interaction, peer tutoring, and cross-grade tutoring.
2. The concepts involved here differ from the one above in terms of the approach to the learning of basic skills. "Self-teaching" refers to both a lot of individualized, self-directive, and independent work and to a lot of group and student interactive work. "Exploration" and "doing" refer to such activities as working on games, puzzles, shopping, cooking, camping, and taking

part in fund raising drives or ecology projects, etc. The point here is that these activities are intrinsically interesting and thereby instrumentally effective in the teaching of basic skills. Finally, such classes generally have students (in addition to the teacher) making many task decisions.

3. An emphasis on self-other awareness means, in other words, an emphasis on affective as contrasted with cognitive learning. Consequently, a lot of activities will be couched in the following concepts: developing one's self-worth, accepting responsibility for one's actions, developing life goal directions, understanding different cultures and thereby learning to accept others, promoting cooperative interchanges, etc. These classrooms will likely emphasize student-student interactions, peer- and cross-grade tutoring, and positive (not negative) feedback.
4. "Cultural diversity" is pretty much self-explanatory, but a classroom that emphasizes this will be involved with cross-cultural comparisons between Anglo-American and the other cultures studied. Cross-cultural comparisons will range over history, food, dance, music, and values, etc. Bilingual classrooms should fall on the high end of this scale.
5. "Fine arts" is also self-explanatory, including such activities as ceramics, crafts, drawing, painting, music, written and verbal expression. Like 2 above, fine arts activities are intrinsically enjoyable, and thereby mediate the learning of basic skills. Throughout all of these activities creative self-expression is promoted.
6. Whether or not a teacher expects a lot of work to get done in class is not as difficult to scale as it may appear. The pilot study shows there is virtually no disagreement on this item, so reference to your own experience should suffice. There is one good indicator, though: Look for negative feedback from the teacher in terms of work not getting done. These comments usually indicate that the class is task-oriented, and teachers of these classes generally expect a lot of work.

7. Diversity of activities is scaled on (1) activities/tasks over and above those involving basic skills; and (2) the number of different instructional modes employed. (1) Since descriptions of all classrooms say they teach basic skills, your point of reference will be a minimum of two activities--language arts and mathematics. See Activities Code and 5b of FIVE MINUTE OBSERVATION for lists of activities/tasks. See 5d of FIVE MINUTE OBSERVATION for a list of different instructional modes. There should be a high correlation between (1) and (2): If there are few activities there will be generally few instructional modes.
8. Diversity of instructional materials refers to diverse materials used, and the existence of diverse learning locations. Items 5a and 5c of FIVE MINUTE OBSERVATION list different locations and materials, respectively. Generally, if there are a lot of different learning locations there will likely be a high diversity of materials used.
9. For "individualized instruction" see the definitions under 5d and 6f of FIVE MINUTE OBSERVATION. The key terms here are "working alone," "working quietly," and "not working interactively."
10. Whether or not students maintain interest in the class can be seen only in terms of student behavior, i.e., whether or not students are participating. See 6c of FIVE MINUTE OBSERVATION for the definition of "participation." The key terms here are "actively involved," "doing," and "acceptable behavior."

Appendix C

ANALYTIC SAMPLE DESIGN FOR THE OBSERVATIONAL STUDY OF DIVERSITY

INTRODUCTION

If we assume a definable domain of relevant variables that may be measured on any unit of interest, induction of a typology of elementary units based on these variables may be seen as the process involved in answering two broad questions:

(1) Is a hierarchical or other grouping of units educible from the data?

(2) If so, which unit of the given hierarchy seems to guide the grouping?

In this appendix, we outline a model of data generation for the observational study undertaken, in an effort to make two points:

(1) The data may be said to reflect ongoing real process to the extent that the observer is not an important factor in generating the data; so we wish to estimate parameters of the model in order to determine the extent to which observers are implicated. This is the main reason for the proposed study and associated model.

(2) Whether we are primarily interested in finding the most parsimonious model that provides a satisfactory fit, or in estimating proportions of variance attributable to the main structural factors, i.e., to classroom, mini-school, and school, we must minimize confounding of factors, especially these three, since the double hierarchy in which they are in fact arranged itself involves inevitable confounding. Hence, crossing of other factors and replication at every level are necessary if confounding is to be minimized. This implies that, in the particular case of classroom, mini-school and school, within-level replication is necessary for the observational study per se.

In the preliminary study described here, deliberate confounding is used as a device to maximize the return from limited resources. However, if this sort of confounding (involving the absence of replication at each hierarchical level) were carried over to the main study, we would be hampered in our efforts to answer question (2). Hence, the main study sample was drawn to ensure full replication at the critical levels. Since full replication was impaired by improper choice of replacement units for unavailable originals in the sample, the main study analytic sample was reduced by the elimination of unreplicated units.

DESIGN PRELIMINARIES

Presumably, the design for data collection is guided by an implicit model that includes those factors thought to be implicated in the generation of the process, the observation of which yields tangible data. We propose, then, to make this model explicit, and thereby suggest a determinate partition of the variance.

Now, while the model per se does not ascribe differential status to the hypothesized generative factors, the fleshed-out conceptual scheme that dictated the model does. This ascription of status implies a rational order for testing the full model against various restricted alternatives. That is, the interpretability of observed effects of certain model factors is dependent on the prior interpretation of others. In the case at hand, these prior factors may be generally classified as conditional or artifactual so that observed effects would, in theory, be taken as indicative of some degree of questionable data integrity. A finding of relatively large observed estimates of effect of these factors would, to varying degrees, inform interpretation of what we may call focal or substantive factors.

Leaving aside for the moment the factor that we might call "subject being taught in the classroom," it seems reasonable to postulate a model of data generation that depends upon six factors, three of which we may classify as conditional and three as substantive. (See Table C.1.)

The identification of the model in terms of these factors is based upon the following procedure of data collection. A sample of classrooms taken from a sample of mini-schools, which in turn are taken from a

Table C.1

FACTORS EXPLICIT IN HYPOTHESIZED GENERATIVE MODEL OF OBSERVATIONS

Name	Symbol		Class[a]	Sample Status[b]		N[c]	
	English	Greek		Local	Model	Symbolic	Example
Observer	O	α	C	R	R	o	4
Period	P	β	C(s)	R	R	p	3
Occasion	T	γ	C(S)	R	R	t	8
Classroom	C	δ	S	F	R	c	5
Mini-school	M	χ	S	F	R	m	6
School	S	φ	S	F	R	s	10

[a]Denotes the theoretical status of a factor:

C = condition of observation; theoretically of no interest, but a possible source of error variance

S = substantive factor; theoretically controls

C(s) = just possibly substantive; variance most likely to be background noise

C(S) = in the present theoretical model, akin to P; potentially substantive

[b]Denotes whether the number of examples of the factor included in the study exhausts the universe of instances (fixed) or is a (theoretically random) sample.

Local = actual sample status in the present study
Model = assumed sample status for analysis and inference

[c]Number of examples of the factor:

Symbolic = notation in the general model
Example = number assumed in numerical example

sample of mini-schools, which in turn are taken from a sample of schools (where any or all of the samples may be exhaustive) are observed by several observers at two periods on each of several occasions, which are separated by a certain number of days. As will be shown below, given that this is an accurate model specification, the model based upon this procedure involves minimal confounding.

As a preliminary to an abstract statement of the full model, we will here, in the interests of increased comprehensibility, briefly comment upon each of these factors in turn.

1. Observer. As suggested above, most of the data upon which the proposed typology will be based are derived from direct observation of the classroom process. Now, although (hopefully) the factor that we are calling "observer" will have no effect upon that process per se, it is quite likely that the observer will have an effect upon the data. That is, the process yields data only through the medium of observer. Hence, a model that purports to describe a generation of data must, clearly, include as one possible factor the observers who collect them. However, equally clearly, we are anxious to reduce to a minimum the amount of noise contributed in the transmission process by the observer. Said differently, to the degree that the data are affected by individual observer, the data will reflect not the process itself, but idiosyncratic interpretations of it. We must then include the observer factor in the model, not so much in the interests of explaining the data maximally as to assess the amount of noise that this factor will contribute. The observer factor, then, is preeminently a representative of the class of factors that we have called artifactual or conditional.
2. Period. Each time a classroom is observed, each observer will complete the entire observational schedule two or more times. Primarily, this is in the interest of assessing the amount of short-term and essentially nonconsequential

variation that occurs in the process of observation. It will necessarily, by increasing the number of data points, tend to make parameter estimates more stable. However, it is possible, if not likely, that regular periodic short-term variations may in fact characterize certain classrooms and/or programs. Should this be the case, we need, of course, to have some measure of the extent and locus of this variation, since the periodicity might in fact be a substantive discriminator of educational units. Primarily, however, the period factor is conceptually conditional or artifactual.

3. Occasion. In contrast with the period factor, this factor is intended to represent the effects of longer-term variation in process. Ordinarily, discrete occasions are separated by several days. Now, since our basic conception of educational units is, so to speak, nonevolutionary, we tend to see the occasion factor as representing simply replicated observations of typical unit process. However, it is conceivable that certain classroom and/or education programs do exhibit regular trends over time. It seems far more likely that these trends would appear in discrete occasion observations than in period observations within occasions. Therefore, while we basically think of the occasion factor as conditional or artifactual, and consequently a contributor to stable estimations in the same way that period is, there is obviously more reason to entertain the hypothesis that occasion may, in fact, be not conditional or artifactual but focal or substantive.

We come now to the factors which, from the present viewpoint, constitute the substantive or focal factors in which we have considerable theoretical, as opposed to merely methodological, interest.

4. Classroom. This is, of course, the traditional basic unit for the organization and delivery of elementary education. In this sense, then, it is the basic unit of observation and of analysis.

5. Mini-school. In the sense of experimentation, this is the factor which we might refer to as the "treatment" factor. In any case, it is the most critical of the focal factors because, in the present context, it is the theoretically new basic unit of organization and delivery of elementary education of which the classroom (in this formulation) is merely a replication. Finally, it is the mini-school which, in the best of all possible worlds, will turn out to be the basic unit of typology construction.
6. School. It is, of course, a commonplace that the individual school (whatever it may consist of) seems to have potent influences upon certain student outcomes, and it is not, therefore, fanciful to suppose that it also has a significant effect upon the sort of process with which we are immediately concerned.

These, then, are the basic factors which, singly and in combination, we hypothesize to constitute the specifications for a model of data generation in this observational study.

We now turn to a full symbolic specification of the model. Inasmuch as the primary focus of the present analysis is upon partitioning the variance in observations, we have, for the sake of clarity, initially shown the fully specified model in a traditional ANOVA layout in Table C.2.

Note first that, as stated above, the design for data collection that is reflected in the model of Table C.2 is confounded to the minimal extent consonant with the reality structure of the system under observation. That is, such confounding as appears in this model does not result from design factors, but from the basically doubly hierarchical nature of the units under observation. That is, the educational structure in Alum Rock is such that individual classrooms are nested within mini-schools, which are in turn nested within schools. The only difference between this and a model based upon what we may call the traditional educational system is the fact of the double nesting, which results from the fact that mini-schools,

Table C.2

FULL ANOVA MODEL WITH ALL FACTORS MINIMALLY CONFOUNDED

Factor	Degrees of Freedom		Confounding[1]
	Symbolic	Example	
[c]Observer (O)	(o-1)	3	
[c]Period (P)	(p-1)	2	
[c]Occasion (T)[2]	(t-1)	7	
[s]Classroom (C)	ms(c-1)	240	CM, CS, CMS
[s]Minischool (M)	s(m-1)	50	MS
[s]School (S)	(s-1)	9	
OP	(o-1)(p-1)	6	
OT	(o-1)(t-1)	21	
OC	(o-1)[ms(c-1)]	720	OCM, OCS, OCMS
OM	(o-1)[s(m-1)]	150	OSM
OS	(o-1)(s-1)	27	
PT	(p-1)(t-1)	14	
PC	(p-1)[ms(c-1)]	480	PCM, PCS, PCMS
PM	(p-1)[s(m-1)]	100	PMS
PS	(p-1)(s-1)	18	-
TC	(t-1)[ms(c-1)]	1680	TCM, TCS, TCMS
TM	(t-1)[s(m-1)]	350	TMS
TS	(t-1)(s-1)	63	
OPT	(o-1)(p-1)(t-1)	42	
OPC	(o-1)(p-1)[ms(c-1)]	1440	OPCM, OPCS, OPCMS
OPM	(o-1)(p-1)[s(m-1)]	300	OPMS
OPS	(o-1)(p-1)(s-1)	54	
OTC	(o-1)(t-1)[ms(c-1)]	5040	OTCM, OTCS, OTCMS
OTM	(o-1)(t-1)[s(m-1)]	1050	OTMS
OTS	(o-1)(t-1)(s-1)	189	
PTC	(p-1)(t-1)[ms(c-1)]	3360	PTCM, PTCS, PTCMS
PTM	(p-1)(t-1)[s(m-1)]	700	PTMS
PTS	(p-1)(t-1)(s-1)	126	
OPTC	(o-1)(p-1)(t-1)[ms(c-1)]	10080	OPTCM, OPTCS, OPTCMS
OPTM	(o-1)(p-1)(t-1)[s(m-1)]	2100	OPTMS
OPTS	(o-1)(p-1)(t-1)(s-1)	378	
Mean	1	1	
TOTAL	optcms	28800	

[-] = Artifactual or Error Component
c = Condition Factor
s = Substantive Factor
1) = Display of components confounded with row terms as a result of the hierarchical structure of the system.
2) = All components involving T would usually be analyzed for trends.

rather than classrooms, have been deliberately structured to be the basic unit of education, and that mini-schools are, by administrative decision, confined to particular school buildings.

Now, a little consideration of the model layout in Table C.2 should demonstrate that, in terms not of tests of significance, but rather in apportionment of proportions of variance, it is a maximally flexible design for answering three conceptually distinct sorts of questions.

The first of these three broad classes of allocations we may call variance attributable to purely artifactual sources, or, more commonly, variance attributable completely to unreliability. Basically, this class is defined as all factors of order higher than one which involves the observer factor. An overall, as well as a more detailed, analysis of the relative and absolute size of this class of components appears in Table C.3.

As a result of the partially hierarchical nature of this design, of the total of 64 definable components, only 32, or 50 percent, are estimable. Exactly half of both the definable and the estimable components involve the observer factor. The basic notion underlying this categorization is quite simple. It says simply that, apart from what we may call an overall difference in leniency (or what we would call, in another language, an observer main effect), observations by individual observers on any other single factor in the design should essentially agree. Put differently, the observer factor should not interact with any other factor in the model; or, perhaps better, components of variance attributable to any factor involving the observer of order higher than one should be essentially zero. Now, a rough first approach to assessing the variance attributable to unreliability is simply the following: Leaving aside the mean, we should simply divide the total variation in the data into a part attributable to the observer main effect, a part attributable to all other factors involving the observer, and a residual that would be attributable to nonobserver factors of all orders. If, then, we found that the second portion was a very small portion of the total variation in the data, we could be reasonably confident that observer unreliability, or error

Table C.3

VARIANCE ATTRIBUTABLE TO ARTIFACT OR UNRELIABILITY

(1) Order	Factors: Definable (2) Total	(3) Arti-factual	(4) Proportion [(3)/(2)]	Estimable (5) Total	(6) Arti-factual	(7) Proportion [(6)/(5)]	Total Proportion [(5)/(2)]
0	1	0	.00	1	0	.00	1.00
1	6	1	.17	6	1	.17	1.00
2	15	5	.33	12	5	.42	.80
3	20	10	.50	10	7	.70	.50
4	15	10	.67	3	3	1.00	.20
5	6	5	.83	0	0	.00	.00
6	1	1	1.00	0	0	.00	.00
Total	64	32	.50	32	16	.50	.50

of measurement, would not be a significant factor in the analysis. However, as will be seen below under a discussion of special reliability studies, dependent upon the degree and patterning of planned confounding in the actual data collection effort, the estimation of additional specific components of variance may be advisable.

The second major category of interest comprises terms involving the period and occasion factors. Of these, we would be particularly interested in those further involving classroom, mini-school, or school factors. Since they would be of no assistance in distinguishing programs, in this context, the main effects of both period and occasion are of little or no interest. However, relatively large and/or reliable proportions of variation attributable to higher order factors would also be potentially interesting, since they might lead to reasonable substantive (as opposed to artifactual) interpretations of these factors.

The third and substantively most interesting class of sources is, of course, composed of the last three: classroom, mini-school, and school. In addition to the fact that it is these factors which are

theoretically essential to this investigation, the relative proportions of variation in the data that are attributable to these three factors may be interpreted as indicators of relative likelihood that the units which are the levels of the respective factors are the strongest candidates for establishing the basis of the typology. To see this, it may be easier to begin with an examination and discussion of Table C.4.

Table C.4

SCHEMATIC ILLUSTRATION OF SUCCESSFUL HIERARCHICAL PARTITIONING OF VARIATION BETWEEN CLASSROOMS

	Hierarchical Stage (DFs)		
Factor	1	2	3
Mean	1	1	1
C	cms-1	--	--
CwM[a]	--	ms(c-1)	ms(c-1)
M	--	ms-1	--
MwS	--	--	s(m-1)
S	--	--	s-1
Total	cms	cms	cms

[a] XwY - X nested within Y; the factor so denoted is residual after the nesting factor.

As the diagram in Table C.4 tries to make clear, in any partially hierarchical design, the total variation and degrees of freedom associated with any term in the design in the hierarchy which is not at the top are partitioned into a proportion attributable to the next hierarchy level and what may be interpreted as a residual portion remaining for attribution to the next lower factor. For example, in terms of Table C.4, leaving aside the mean, at the first stage in the hierarchy (i.e., in which we are dealing with, so to speak, a single factor hierarchy), all of the variation that lies between classrooms is directly and completely attributable to the classroom factor. However, with the

addition at stage two of the mini-school factor, the total variation at stage two is apportioned into a part lying wholly between mini-schools, and a residual portion (the only portion now attributable directly to classroom) which lies wholly between classrooms, but only within mini-schools. With the addition at stage three of the school factor, as we see in Table C.4, while the portion attributable to the residual between classroom within mini-school variation remains the same from stage two to stage three, the proportion attributable at stage two to the mini-school itself becomes divided into a portion lying wholly within schools, and a residual attributable now only to mini-schools within schools.

Suppose now that at, say, stage two, the amount of variation attributable to the classroom factor at stage one equals the amount attributable to the classroom residual at stage two. The clear implication, of course, would be that in terms of these data there was no distinction between mini-schools as such, except insofar as there might be distinctions between a subset of classrooms in one mini-school and a subset in another. In short, under these circumstances it would be clear that, since the data would not discriminate one mini-school from another, the basic unit upon which a typology must be based would perforce be not the mini-school, but the classroom. Of course, this is overdrawn for emphasis, but the general thrust of the argument is clear. Given the complete three-stage hierarchy, the relative proportion of the variation attributable to the three hierarchical factors should be reasonable indicators of the likelihood that these factors will be found to underlie a reasonably discriminating typology.

However, since there will almost surely be a substantial number of degrees of freedom attributable to each of the factors even at the highest hierarchical stage, unless the relative proportion of variation attributable to one of the factors is extremely small, it would not be particularly wise to stop with the overall assessment. That is, using a procedure analogous to the definition of specific comparisons or contrasts in the traditional experimental analysis, we almost surely would want to investigate the extent to which, for any of these factors, some higher order grouping of levels (but with degrees of freedom less

than the total attributable to the factor) might be relatively explanatory. Finally, we will briefly mention the problem of confounding within the hierarchy. That is, to take an extreme case, if a certain school was represented by only one mini-school which in turn was represented by only one classroom, it is obvious that--in terms of actual estimation--the only level which could possibly contribute to a typology would be the highest level, which in this case would be the school. However, we will go into this subject a bit more thoroughly in our discussion of the model underlying the special reliability study.

Finally, Table C.5 presents the structural model corresponding to the data collection design with which this paper was introduced. In examining the model, however, it is necessary to bear in mind that, as shown in detail in Table C.2, not all of the factors included in the structural model can be estimated free of confounding. In one sense, absent the assumption that all definable factors not explicitly represented in the structural model have effects equal to zero, at least those factors explicitly represented in the structural model which are confounded with one or more which are not represented will yield uninterpretable estimates. (For an explicit description of this phenomenon, see Table C.7.) However, as will be seen below, from the viewpoint of estimation of crucial effects, this theoretical bias will turn out to be much less serious than might at first appear.

Now, depending upon the assumptions made about the sampling status of the various factors, it is straightforward, on the basis of the structural model, to estimate any components of variance of interest. But in spite of the fact that apparently reasonable assessments of sampling status for all of these factors appear in Table C.1, it is still true that there is certainly plenty of reasonable ground for disagreement. Given this situation, which implies a very large number of alternative specifications of sampling status, it will probably serve no useful purpose at this stage to try to write expected mean squares for all of the factors in this model. However, this is a very straightforward (if laborious) task.

Table C.5

FULL STRUCTURAL MODEL

$$
\begin{aligned}
X_{optc} = \mu
&+ \alpha_o + \beta_p + \gamma_t + \delta_{c(ms)} + \chi_{m(s)} + \phi_s \\
&+ \alpha\beta_{op} + \alpha\gamma_{ot} + \alpha\delta_{oc(ms)} + \alpha\chi_{om(s)} \\
&+ \alpha\phi_{os} + \beta\gamma_{pt} + \beta\delta_{pc(ms)} + \beta\chi_{pm(s)} \\
&+ \beta\phi_{ps} + \gamma\delta_{tc(ms)} + \gamma\chi_{tm(s)} + \gamma\phi_{ts} \\
&+ \alpha\beta\gamma_{opt} + \alpha\beta\delta_{opc(ms)} + \alpha\beta\chi_{opm(s)} \\
&+ \alpha\beta\phi_{ops} + \alpha\gamma\delta_{otc(ms)} + \alpha\gamma\chi_{otm(s)} \\
&+ \alpha\gamma\phi_{ots} + \beta\gamma\delta_{ptc(ms)} + \beta\gamma\chi_{ptm(s)} \\
&+ \beta\gamma\phi_{pts} + \alpha\beta\gamma\delta_{optc(ms)} + \alpha\beta\gamma\chi_{optm(s)} \\
&+ \alpha\beta\gamma\phi_{opts}
\end{aligned}
$$

Notation:

Greek letters denote parameters of the model.
English letters as subscripts denote number of factor levels and factor relationships, as in the examples which follow:

Term	Relationship
$\beta\gamma_{pt}$	PERIOD is crossed with OCCASION.
$\delta_{c(ms)}$	CLASSROOM is nested within MINISCHOOL and SCHOOL.
$\chi_{m(s)}$	MINISCHOOL is nested within SCHOOL.
$\beta\chi_{pm(s)}$	MINISCHOOL (within SCHOOL) is crossed with PERIOD.

In Chap. 1 we stressed that the model we would discuss, assuming that it was correctly specified (i.e., that we had, roughly, identified all relevant factors), would allow us to assess effects that were minimally confounded (i.e., such that the estimates were uncontaminated with the effects of other variables to the extent permitted by the structure of the system which we were observing). We have also noted, however, that a large proportion of the estimates which this model would permit us to make were in fact of no substantive interest, but rather designed to provide a degree of assurance that inferences which we might make about effects of substantive interest were not unduly contaminated by artifactual influences. In that sense, then, the freedom which is allowed us by the fully specified model in Table C.2 is a good deal more freedom than we want or require.

To put this in perspective, suppose that, from some reliable source, we knew in advance of our study that all of the parameters specified in the full model in Table C.2, which involved observer at any order, were simultaneously zero. Clearly, if this were the case, we would simply exclude all of these terms from our model, since a model in which they were excluded would be as perfectly specified as one in which they were included; we would eliminate thereby, as Table C.3 shows, roughly 50 percent of the factors, but in point of fact a much larger proportion than that of parameters. To see this, consider that, under the numerical specifications which we have attached as an example to Table C.2, each observer would be making a total of 7,200 observations. If, as we supposed above, we knew that the observer effect was zero--or, said differently, that one observer was entirely equivalent to another under all conditions--it would clearly make no difference which observer we used for which observations.

At this stage, of course, we really know nothing at all about the size of observer effects. However, it should be possible to determine or estimate the size of these effects on the basis of considerably fewer observations than 7,200. That is, in the interests of efficiency, it would seem most sensible that, as a preliminary study, we set out simply to estimate the sizes of the effects involving observers, which, if carefully done, should provide us with estimates sufficiently

precise to allow us to project them into a full study. Therefore, it seems most sensible to perform a preliminary reliability study, and as we will see below, it is possible to do this with a great deal less wasted effort and resources than the effort implied by a fielding of the design set out in Table C.2. The form that this might most reasonably take we can consider to be a version of the model in Table C.2, in which we estimate considerably fewer parameters. Table C.6 presents an outline of such a reduced model, again set out in an ANOVA design.

Aside from the fact the total number of degrees of freedom in the second model is less than one percent of the number in the first model proposed, the most blatant difference between the two is the presence of a considerably increased amount of confounding. However, as closer examination will show, this impression is to some extent illusory. In fact, a comparison of the two models shows that, for every confounded factor in the second, there was a total of three estimable factors in the first, of which only one was itself confounded. This results from the fact that in the second model, the classroom, mini-school, and school factors are by design completely confounded.

The logic of this procedure is extremely simple. However, it proceeds from the fact that, in the end, a random sample of classrooms is automatically a random sample of both schools and mini-schools. This being the case, the simplest way to get a full representation, for any given number of observations of classrooms, of both schools and mini-schools, is to ensure that the classroom be sampled randomly, subject to the condition that each classroom sample produce also a different mini-school and school. Since we do not know which of these factors is the most explanatory (or equivalently, the most likely unit underlying a typology) it is to our advantage to get, for purposes of reliability assessment, the fullest possible representation of all categories. That, in essence, is precisely what this design demands.

Table C.7 presents the structural model for the ANOVA layout for the preliminary study of observer reliability or observer agreement. This model has fewer terms than the model of Table C.5. This is clearly a result of the deliberate decision completely to confound, in the reliability study, the effects of School and Mini-school with those of Classroom.

In order to make clear, however, that changes in models do not dictate changes in reality, the model in Table C.7 uses the symbol δ' instead of δ, the general symbol for Classroom parameters used in Table C.5. Moreover, all parameters of the reliability model involving δ' have been written in expanded form in the lower part of Table C.7 in terms of the parameters of the model of Table C.5. That is, estimates of any parameters in the reliability model containing the symbol δ' cannot be unambiguously interpreted. Estimates of these parameters may result from any linear function of the appropriate parameters explicitly represented in the general mode of Table C.5. However, since the latter are not separately estimable in the reliability model, it is impossible to estimate the parameters of these functions from these data.

This is, of course, also true of certain parameters in the general model in Table C.5, as can be seen by comparing the structural model in Table C.5 with the equivalent ANOVA layout of Table C.2. However, the confounding in the former model is a reflection of the structure of the system being modeled. In the latter, the degree of confounding has been deliberately increased for reasons of inferential economy. This distinction, of course, has no effect upon estimability, but it is nonetheless important: Parameterizations of models that render unestimable functions that the designer has no interest to estimate can often (as here) effect significant economies of effort and resources.

From a reliability point of view, the most crucial estimates to be made are those involving both observer and classroom as a factor, since our substantive interests lie precisely in the classroom/mini-school/school unit. Therefore, for a given total number of degrees of freedom, the requirement of efficient design dictates that the degrees of freedom allocated to these factors, relative to all others, should be maximized. A glance at the design sketch in Table C.6 shows that 180 degrees of freedom, or 75 percent of the total, are allocated to just these factors.

Moreover, to a large extent, it is just this allocation of degrees of freedom that minimizes the effects on our inferences of the degree

Table C.6

ANOVA MODEL FOR PRELIMINARY RELIABILITY ASSESSMENT

Factor	Preliminary Reliability Assessment DF[a]		Confounding[b]
Observer (O)	(o-1)	5	
Period (P)	(p-1)	1	
Occasion (T)	(t-1)	1	
Classroom (C)	(c-1)	9	M, S, MS, CM, CS, CMS
OP	(o-1)(p-1)	5	
OT	(o-1)(t-1)	5	
OC	(o-1)(c-1)	45	OM, OS, OMS, OCM, OCS, OCMS
PT	(p-1)(t-1)	1	
PC	(p-1)(c-1)	9	PM, PS, PMS, PCM, PCS, PCMS
TC	(t-1)(c-1)	9	TM, TS, TMS, TCM, TCS, TCMS
OPT	(o-1)(p-1)(t-1)	5	
OPC	(o-1)(p-1)(t-1)	45	OPM, OPS, OPMS, OPCS, OPCM, OPCMS
OTC	(o-1)(t-1)(c-1)	45	OTM, OTS, OTMS, OTCS, OTCM, OTCMS
PTC	(p-1)(t-1)(c-1)	9	PTM, PTS, PTMS, PTCS, PTCM, PTCMS
OPTC	(o-1)(p-1)(t-1)(c-1)	45	OPTM, OPTS, OPTMS, OPTCS, OPTCM, OTPCMS
Mean		1	
Total	optc	240	

[a]Based on o=6, p=t=2, c=10.

[b]Components confounded with row terms.

Double underlining identifies components confounded with row terms in full model.

Single underlining identifies components not confounded with row terms, but mutually confounded in full model.

No underlining identifies components unconfounded in full model.

Table C.7

STRUCTURAL MODEL FOR RELIABILITY DESIGN

$$
\begin{aligned}
Y_{optc} = {} & \mu \\
& + \alpha_o + \beta_p + \gamma_t + \delta'_c \\
& + \alpha\beta_{op} + \alpha\gamma_{ot} + \alpha\delta'_{oc} + \beta\gamma_{pt} \\
& + \beta\delta'_{pc} + \gamma\delta'_{tc} \\
& + \alpha\beta\gamma_{opt} + \alpha\beta\delta'_{opc} + \alpha\gamma\delta'_{otc} \\
& + \beta\gamma\delta'_{ptc} + \alpha\beta\gamma\delta'_{optc}
\end{aligned}
$$

* * * * * * * * * * * * * *

$$\delta'_c = [\delta_{c(ms)} + \chi_{m(s)} + \phi_s]$$

$$\alpha\delta'_{oc} = [\alpha\delta_{oc(ms)} + \alpha\chi_{om(s)} + \alpha\phi_{os}]$$

$$\beta\delta'_{pc} = [\beta\delta_{pc(ms)} + \beta\chi_{pm(s)} + \beta\phi_{ps}]$$

$$\gamma\delta'_{tc} = [\gamma\delta_{tc(ms)} + \gamma\chi_{tm(s)} + \gamma\phi_{ts}]$$

$$\alpha\beta\delta'_{opc} = [\alpha\beta\delta_{opc(ms)} + \alpha\beta\chi_{opm(s)} + \alpha\beta\phi_{ops}]$$

$$\alpha\gamma\delta'_{otc} = [\alpha\gamma\delta_{otc(ms)} + \alpha\gamma\chi_{otm(s)} + \alpha\gamma\phi_{ots}]$$

$$\beta\gamma\delta'_{ptc} = [\beta\gamma\delta_{ptc(ms)} + \beta\gamma\chi_{ptm(s)} + \beta\gamma\phi_{pts}]$$

$$\alpha\beta\gamma\delta'_{optc} = [\alpha\beta\gamma\delta_{optc(ms)} + \alpha\beta\gamma\chi_{optm(s)} + \alpha\beta\gamma\phi_{opts}]$$

of confounding built into this design. For example, if the relative size of the component of variance attaching to one of the observer/classroom factors, or, alternatively, the relative proportion of the total variance in the system, is quite small, then the confounding is of no consequence. The conclusion would be that, whatever the source of the effects, it was in any case small. If one or more of these factors should produce distressingly large effects, however, we will be able to pinpoint the source of the effect without any considerable loss of inferential freedom, and, on the basis of that exercise, to take steps to correct it.

To be more specific, consider the observer by classroom interaction factor in Table C.6, which has associated with it 45 degrees of freedom. In the estimations incident to testing the model set out, we are not only partitioning the variance but also, in the case of a particular factor in the discussion, estimating 45 parameters, one for each degree of freedom. Then, if the influence of this factor as a whole is found to be large, it will be quite simple, by examining the individual parameter estimates, to discover which of the individual effects are responsible for the size of the overall effect that we observe.

For example, it might well turn out that the size of the effect resulted from the unreliability of some subset of observers. In this case, of course, we would attempt to deal with the problem on that level. A little additional analysis might also reveal the effect to be the result of increasing skill over time. In that case, the indicated action might be to give the observers more training before executing the study. Of course, it is always possible that we would find these effects to be specific to some subset of classrooms. In this case, we would face a problem self-imposed by the confounding we specified in the design. That is, given the sampling design specified, it would not be clear whether the problem had to do with a classroom, a mini-school, a school, or some combination of these. In this case, the most reasonable course would be to appeal to sources of information outside our data. However, this seems an unlikely occurrence. That is, the most reasonable and straightforward interpretation of

such a finding, in the absence of unknown unusual factors specific to a particular observation that might explain the finding, would be a hypothesis that the behavior in this classroom was sufficiently ambiguous, as compared to all other classrooms, to make decision on the part of observers extremely chancy. On a priori grounds, this seems highly unlikely, which is precisely the sort of prior assessment that led to the decision to confound. In the unlikely event that this does occur, it will probably be possible to trace the source and make decisions on that basis.

At this point, we should briefly note the considerations underlying the decision to conduct a preliminary reliability study. We need not expand on the matter of wasted resources, which we have referred to above. The second major consideration, however, is considerably more weighty, and amounts to the following: If we conduct the entire study as a combination effort to assess the effects of both artifact and substance, and the artifactual effects prove large, in the limit we will have wasted the entire study--in particular those parts that were devoted to an investigation of substantive effects. If we devote a much smaller preliminary study almost entirely to an assessment of artifact, however, and find large effects, we will have wasted little or nothing on a substantive study. We will also, given a proper design, be able to pinpoint the difficulties, and, in general, remedy them before mounting the substantive study.

Hence, for the next round of observations, we will collect our data in two stages. The first will be a reliability study very much like that schematically described in Table C.6, while in the second (given that we carefully prosecute, analyze, and act upon the results of the first) we can largely ignore the problems posed by the existence of artifactual effects. In the limit, we will need to have little concern with specific assignments of observers. In general, given that observer effects are found to be nonzero, it is wise to randomize them throughout the entire observation design, thereby achieving what we may think of as an even spread of observer error.

We next turn to the topic which we may very generally call the problem of missing data. In the present case, this generally takes the

form of extremely poor sampling. Implicit in our discussion (of a preliminary reliability study) was the inference that, given total confounding of three substantive factors in the design, while we could observe effects attributable to these factors, we could in fact tie them explicitly to none. Now, while this is perfectly acceptable for a study that is only incidentally concerned with substantive effects, it is at least serious and possibly fatal to a study in which the emphasis is reversed. To set the discussion in a concrete context, we will consider the results of the sampling that produced the observational data for the spring 1975 observational study. Table C.8 presents an outline of the resultant sample, in terms of school, mini-school, and classrooms.

As Table C.8 shows, we have 51 classrooms that produced reasonably usable data, representing 22 mini-schools and 10 schools. Because one of the primary substantive questions of the present study is the relative allocation of variation in the data between school, mini-school, and classroom, any sampling design that makes this allocation difficult or impossible is clearly unacceptable for our purposes. Now, given a reasonable distribution of the 51 classrooms, we would expect to have 10 schools and 22 mini-schools represented in a usable way. However, a cursory examination of the distribution of classrooms across mini-schools and schools in Table C.8 shows that a considerable proportion of the data is useless. Compare, for example, the schools labeled E and G. We can clearly assess the amount of variation between the two, but we cannot possibly explain it. We have no way of knowing if the variation was due to factors associated with the school, mini-school, or classroom. In short, for our purposes, the data from schools E and G are useless and must be discarded. Table C.8 reveals other examples of this problem. For example, while the variation within school D can be assessed, an attribution of this variation to mini-schools or classrooms is impossible. While from some points of view this is better than the situation with respect to schools E and G, from the point of view of our overall study, the difference is strictly minimal. In Table C.8, those schools and mini-schools generating unconfounded data are marked with a superscripted plus sign. In Table C.9, we illustrate, in terms of

Table C.8

EXAMPLE OF PARTIAL CONFOUNDING IN C/M/S HIERARCHY
SPRING 1975 DATA

School	$M_w S^a$	N_{cwm} [b]	School	$M_w S$	N_{cwm}
A^+	a^+	3	F^+	m^+	7
	b^+	3		n^+	3
	c	1	G	o	1
	d	1	H	p^+	2
B	e	3		q^+	3
	f	1		r^+	3
C^+	g^+	4	I	s	1
	h^+	3		t	1
	i^+	5	J	u	2
D	j	1		v	1
	k	1			
E	1	1	N= 10	22	51

NOTE: [a]Each lower case letter denotes one mini-school within the school denoted by an upper case letter.

[b]Number of classrooms within each mini-school.

+ = unconfounded.

our substantive factors, the degree and locus of data loss, and consequential inferential infirmity, which results from the confounding obvious in Table C.8. Table C.9 reveals that the loss in terms of the residual classroom factor was very small--about 10 percent. However, it is equally clear that the 50-percent loss of mini-schools and the two-thirds loss of schools renders the entire data base of extremely doubtful utility. This is largely a consequence of the logical structure of partially hierarchical designs. That is, the effects of irrational sampling are far more likely to be felt at the top of the hierarchy

Table C.9

EXAMPLE OF LOSS OF DEGREES OF FREEDOM FROM CONFOUNDING IN SAMPLE SPRING 1975 DATA

Factor	Degrees of Freedom		
	Potential (P)	Usable (U)	Loss [(P-U)/P]
Mean	1	1	.00
S	9	3	.67
$M_w S$	12	6	.50
$C_w M$	29	26	.10
Total	51	36	.29

than at the bottom. Without belaboring the notion, we should simply point out that, when the sample is drawn, we must be extremely careful that all of the factors in which we have substantive interest are well represented, and that the degree of confounding involving these factors is minimal.

We now turn to the problem of the number of occasions upon which a given classroom should be observed. In past research, our tendency has simply been to choose a convenient number and let it go at that. We now know that it is unwise to do so. The choice of the number of these occasions (and, to a smaller extent, their timing or spacing) is deducible from, and must be based upon, a consideration of the sort of process that we believe to underlie the data we will gather.

The guiding principle here is, very briefly, that the structure of the observations must allow us to detect the broadest class of effects that we believe may be significant. For example, we may believe that the classroom process differs in basic ways according to the subject matter being taught, as between the teaching of reading and the teaching of mathematics. We may also independently suppose that the classroom process is either (a) an example of a relatively steady state, or

(b) a process that shows evolutionary longitudinal trends. The case of differences in subject matter is simple to resolve. If we believe they exist, and we are interested in them, we should ensure that our observations are evenly distributed across two areas, other things being equal.

In the second case, if we believe we are simply observing a steady-state system, which implies that one sample is equivalent to any other within measurement error, the number of observations obviously must be equal to or greater than two. If we believe that the process may be evolutionary, however, it is equally clear that the number of observations must be equal to or greater than three and that, in addition, they should be reasonably spread out over time.

Another issue has to do with the typicality of any particular sample. For example, it is fairly obvious that, whether a program exhibits a steady state or an evolutionary trend, the school year will have certain predictable periods during which no observation will be useful in assessing the process in any real sense. To take an obvious example, it is probably absurd to expect any classroom to exhibit its typical behavior in the neighborhood of an important outside event, such as a riot, a holiday, or an epidemic.

Otherwise, in order to allow for the greatest degree of generality, it would appear that the number of observations of each sort of subject matter should be at least three and possibly more. This suggests, at its simplest, a series of at least three different occasions for each classroom, of a duration such that a reasonable sampling can be obtained for the two subject matter areas in question.

If the notion of evolutionary process is to be taken seriously, we must bear in mind the possibility, other things equal, that the evolutionary trend is likely to be more visible earlier in the year than later. On the other hand, we would not like to be placed in the position of mistaking a dampening of random variation toward steady routine (also more visible early in the school year) with a true evolutionary trend. At any rate, these fundamentally different processes are likely to be more easily confused during the early parts of the school years, when the program is shaking down. If we are to take these possibilities seriously, it

would probably be the part of wisdom to allow the number of occasions to be a minimum of, say, six, spread fairly evenly over the year but avoiding special times such as holiday periods.

In reality, that is probably far too much to expect. If the number of observations or occasions is reasonably restricted to only two, say, it would seem wise to make both observations during the spring term following the post-holiday shakedown. Although it seems only remotely likely that this plan will allow us to separate a steady state from an evolutionary process, we would at least obtain the minimum number of occasions necessary for assessing the amount of day-to-day variation in what we would be implicitly assuming to be a steady-state system.

We have set out at considerable length in Chap. 2 above the distinction between a descriptive and an inferential or effective typology, and related it to the difficulty of ascribing effects to individual mini-schools that result from their nesting within schools. Our conclusion was that, from the point of view of effects estimation, the ideal outcome of a typology study would be a classification of mini-schools that was minimally confounded with schools. That is, in the limit, we would like to have a typology which is, in the language of design, fully crossed with schools. The extent to which this objective can be attained clearly depends, however, upon the distribution of the number of mini-schools across schools and the number of classifications of mini-schools, if indeed the mini-school is the proper unit that emerges from the typology study. To illustrate, in Table C.10 is a symbolic layout of the effects of introducing such a crossed typology.

To begin, it is obvious from Table C.10 that introducing a typology factor to cross schools creates additional problems for the already complex partially "nested" hierarchical design. Specifically, the degrees of freedom associated with the mini-school within school residual becomes the sole source of degrees of freedom for both the typology main effect and the interaction of that factor with schools. More important for our present purposes is the fact (also illustrated in Table C.10) that within any particular school, where M is the number of mini-schools within a particular school and G is the number of

Table C.10

SYMBOLIC LAYOUT OF THE EFFECTS OF INTRODUCING
A MINI-SCHOOL TYPOLOGY THAT CROSSES SCHOOLS

Factor	Partition of DF	
	Before	After
Mean	1	1
School (S)	(s-1)	(s-1)
Typology (G)		(g-1)
Mini-school (M)	s(m-1)	s(m-g)
SxG		(s-1)(g-1)
Classroom (C)	ms(c-1)	ms(c-1)
Total	cms	cms

classifications in the typology, the number of degrees of freedom for the mini-school residual in the hierarchical design is equal to (M-1) but in a design incorporating a cross-typological factor, it is (M-G).

It follows, then, that for G that is at all large--say, three or more--and given what is known about the typical number of mini-schools in a particular school, the achievement of total lack of confounding is highly improbable. To take a simple example, imagine that M equals G equals four for every school in the demonstration. It is then clear that the difference between types within any particular school is totally confounded with the difference between mini-schools within any particular school. On the other hand for example, the fact that in the hierarchical design most of the variance is found to lie between mini-schools would lead to the conclusion that, of the three hierarchical units, this was the one most likely to underlie any clustering. It would perhaps be reasonable to assume that the likelihood of residual variation due to mini-schools within types was quite small and that therefore we could simply ignore it. In that case, difficulty would arise only if G were greater than M for some particular school. In that case, however, it would seem most reasonable to entertain the hypothesis

that the type of mini-school was a function of the school itself to some presumably estimable extent. Although that would not solve our inferential problem with regard to educational effects, it should at least make them simpler to understand. Finally, we must consider whether the analyses described in this report are to be univariate or multivariate in form, and what the dependent variable or variables shall be. To take the second problem first, we will simply state that, barring extremely unforeseen circumstances, we would expect the dependent variables to emerge fairly clearly from an assessment, based upon the fairly strong notions which underlay the construction of the instrument, of the covariance structure of the data points. To put it as grossly as possible, it seems most likely that the dependent variables will consist in what we will refer to with some imprecision as a set of scales.

We can now see that if we define our scale in advance of our assessment of reliability, we will involve ourselves in an endless cycle--reinforcing the conclusion that a well-designed, well-executed, and comprehensive reliability study is an absolute necessity.

Appendix D

DERIVED SCALES FROM CLASSROOM OBSERVATION AND TEACHER QUESTIONNAIRE

Several series of summary questions (for observers), and two sets of opinion items (for teachers), were included in the two classroom instruments, in an effort to get at characteristics more molar than the five-minute observation items and the structural questions in teacher self-reports seemed likely to provide.

We did not intend, however, that each question should stand alone. Rather, the notion was that certain linear combinations of them might be derived such that these linear functions would describe individual classrooms in a conceptually coherent way. As an example, we thought that from the summary scales (see the observation instrument) we should be able to find either one bipolar or two reasonably orthogonal dimensions, which we could roughly describe as characterizing classrooms along a dimension of diversity and student-oriented versus traditional.

This appendix communicates the results of that scaling exercise for each of the judgmental item sets included in the instrumentation. To anticipate a bit, the dimensions that emerged from the observation summary items were both strong and coherent, and virtually orthogonal. These were subsequently used as criteria against which the scales derived from other sections of the observation instrument could be mastered.

As a final introductory note, the scales themselves are composed of equally weighted items. It is clear that, if the scales are coherent, alpha estimates will approximate maximum alpha from unequal weighting. If they are not coherent, any sort of weighting is gratuitous.

We begin with the scales from the observation instrument, and then present scales from the teacher questionnaire. All loadings are positive unless a minus sign precedes the item number. The number itself is from the appropriate instrument section.

OBSERVATION INSTRUMENT

Diversity Oriented (items from Summary Scales)

2 = There is an apparent in-class emphasis on self-teaching by discovery, exploration, doing.

3 = There is an apparent in-class emphasis on self-other awareness.

4 = There is an apparent in-class emphasis on cultural diversity.

5 = There is an apparent in-class emphasis on fine arts.

7 = Students engage in a diversity of activities during the day.

8 = Students work with a diversity of instructional materials during the day.

9 = Student learning activities are individualized.

Basic Oriented (items from Summary Scales)

1 = There is an apparent in-class emphasis on basic skill learning.

6 = The teacher seems to expect a lot of work to get done in class.

10 = Students maintain interest in the class.

Open Classroom (items from Classroom Log: Teaching Style

1 = The teacher's questions are open-ended.

2 = The teacher makes use of student ideas in class (restates them, elaborates them, uses them in examples, asks further questions, etc.).

8 = There is discussion of feelings, behaviors, beliefs, and attitudes.

10 = The class works by discussion or other group-interactive processes.

Traditional (items from Classroom Log: Teaching Style

3 = The teacher employs direct (traditional) instructional methods (lecturing, demonstrating, and other explicit instructional techniques) with groups of students

-4 = The teacher's communication focus is individual students.

5 = The teacher's communication focus is groups of students.

14 = Curriculum is divided into subjects which are taught separately (vs. integrated curriculum).

Administrative (items from Classroom Log: Teaching Style)

6 = The teacher spends time interacting with other adults.

-13 = The teacher spends his/her time interacting with students.

18 = The teacher is engaged in administrative, grading, or other non-interactive tasks.

19 = Other classroom adults are engaged in administrative, grading, or other non-interactive tasks.

Academic (items from Classroom Log: Teaching Style)

7 = The teacher provides students with feedback about their academic work.

9 = The class is working quietly, individually.

-11 = The class gives attention to alternative cultural themes.

-12 = The class gives attention to fine arts.

-15 = Team teaching is employed by teachers and/or aides.

Organization (items from Classroom Log: Management Style)

8 = Order is maintained in the classroom.

9 = There is a smooth, efficient transition from activity to activity.

-10 = The teacher gives negative feedback.

-11 = The teacher gives positive feedback.

(N.B.: other items in this set failed to comprise scales.)

Student Active (items from Classroom Log: Students)

1 = Students initiate questions, exchanges, discussions.

3 = Students are involved in self-instruction.

4 = Students select their own activities.

5 = Students select own groups.

12 = Students engage in self-criticism or self-evaluation.

Student Passive (items from Classroom Log: Students)

2 = Students are in the respondent role (answering questions, speaking when called on, etc.).

6 = Students are expected to remain quiet unless called on to speak.

7 = Students are expected to stay where they are unless given specific permission to move about.

8 = Students stay in one place.

11 = The class is task-oriented.

TEACHER QUESTIONNAIRE

Subject Centered (items from question 1, Classroom Description)

a = Children work under adult supervision.

d = Emphasis is on subject matter.

f = All the class is engaged in the same activity.

-k = Children do not interact freely with each other.

m = Children have assigned seats.

Student Centered (items from question 1, Classroom Description)

b = Children gather information on their own.

c = Children initiate interactions with adults.

e = Undesired behavior is ignored.

g = Lesson plan is flexible.

h = Work and play are not distinguished.

Traditional/Academic (items from Opinions)

-30 = Teachers should be equally concerned with social-emotional development and with intellectural development.

-31 = Parents should be encouraged to participate in class activities.

40 = The use of games, toys, and similar equipment and materials should be restricted to free play periods.

41 = Children should be encouraged to ask the teacher's permission before beginning a new activity on their own.

49 = Standardized test scores seem to be good indicators of children's abilities.

Stereotypes (items from Opinions)

32 = Boys generally have more trouble learning language skills than girls do.

38 = Boys are usually more disruptive in the classroom than girls are.

39 = During class hours, there should be more communication between the children and the teacher than among the children.

42 = Girls usually have more trouble learning number skills than boys do.

Discipline (items from Opinions)

35 = Children appreciate firm discipline.

36 = Children learn best when the classroom is quiet.

43 = Children's language errors should be corrected immediately, if possible.

47 = There are many children for whom an informal classroom approach is not suitable.

-48 = Children should not receive grades on their work.

-50 = Anglo students benefit from being able to go to school with minority students.

Ethnic Beliefs (items from Opinions)

34 = Children will make more intellectual and personal-social progress if the teacher is fro' the same ethnic group as the majority of the class.

51 = Bilingual children learn more effectively when instruction is conducted in their first language.

These constructed scales will receive further examination before inclusion in the final analysis. We believe that the DIVERSITY- and BASIC-oriented scales are fairly strong, and correlations among them and the other observation scales are generally in the right direction. The "Teaching Style" group is not entirely clear; but, as sometimes happens, that may be in part because we have not hit on the mot juste for describing them. In general, the "Management" items were very weak; we are skeptical about the one combination that emerged. On

the other hand, the "Students" scales are in general congruent with the two strong "Summary Scales" dimensions, so that one set reinforces the other. From the Teacher Questionnaire, the classroom descriptive scales are, like the "Summary Scales," pretty clear; the opinion items, like the "Teaching Style" items, are weak. An examination of cross-instrument correlations is expected to clarify relationships among derived scales and between them and other observation items.

Appendix E

PROGRAMS EMPLOYED IN ANALYSES

The following are very brief descriptions of the computer programs used to conduct the analyses. Source listings and additional documentation may be obtained from the authors listed; the specially modified versions of these programs may be obtained from The Rand Computation Center.

POLYCON

POLYCON is a versatile program written basically to do multidimensional scaling; since it is a comparatively recent program, it embodies most of the improvements made in such programs since the beginning, based upon past experience with them.

Basically, POLYCON takes as input a matrix of proximity estimates for a set of objects, and attempts to derive a spatial representation of them, such that the order of magnitude of the proximities is reproduced as closely as possible in a given dimensionality.

It permits the user to apply a wide variety of measurement models to the data, and will handle replicated data with ease; e.g., it permits scaling of several independent proximity matrices, perhaps obtained from several different sources, pertaining to the same set of objects.

Like all MDS programs, POLYCON uses an iterative fitting procedure; also like all MDS programs, the cost rises steeply with the number of objects scaled. At Rand, POLYCON runs in 372K. To derive a 3D configuration of 40 points, TIME = 254.92 CPU seconds; for a 3D configuration of 74 points, TIME = 856.81 CPU seconds.

The present version has been modified to handle up to 90 points (from the original limitation of 60); if remodified to the original limitation, POLYCON will probably run in less than 240K.

The documentation is readily available from:

Professor Forrest W. Young
The L. L. Thurstone Psychometric Laboratory
University of North Carolina
Chapel Hill, N. C. 27514

It includes a long theoretical paper on the measurement models included in POLYCON, a very clear user's manual (as these things go), and a very extensive discussion of various options in POLYCON, as well as a detailed discussion of a long, fully worked analysis of a couple of typical datasets.

POLYCON uses a freeform, keyword control card format, which makes it quite easy to use. Known bug: The capability described in the manual to rotate a configuration to a target matrix does not work.

POLYCON was written by F. W. Young and modified by C. Frost.

PROFIT

PROFIT is a program written to assist in the interpretation of spatial configurations of points, usually representing discrete objects, usually derived from a multidimensional scaling program.

Given an N by K matrix of coordinates of points in a space, and a P by N matrix of (usually) independently obtained measures on some property or scale, PROFIT most usefully fits P vectors in the K-space such that the correlation between the original scale values and the projections of the points on the vectors is maximal.

The output includes input echo, computational details, direction cosines for vectors, cosines of angles between vectors, plots of given vs obtained scale values, and a plot of the first two dimensions, showing points and termini of vector projections.

It is easy to change the program limitations (on the number of points, number of properties, so on), and the first part of the source gives detailed directions for doing this. They are accurate: We have two versions operational, differing in the number of points (N) accommodated.

Time and Region: Fitting 10 properties to a 74 x 3 configuration took 11.78 CPU seconds, REGION = 124K.

PROFIT is low-cost and very easy to use (typically, six control cards are required, in keyword format, plus labels).

The main drawback is the plotting restriction to the first two dimensions. To overcome this, rerun PROFIT for a particular dataset,

changing the CONFIGURATION INPUT FORMAT to read the configuration in a different order. For example, to get a plot of first by third dimension in a K=3 configuration, read the configuration using, say (F10.5, 10X,F10.5,T11,F10.5).

Since dimensions (1) and (3) are the first two read, they will be the first two plotted.

PROFIT was originally written by Carroll and Change at Bell Laboratories, modified at the University of Edinburgh, and further modified by C. Frost.

PROCROT

PROCROT is a program that takes a matrix, say N by K, as a target, and allows the user to rotate another matrix, say, N by M ($M \leq K$), to the original matrix by an orthogonal transformation to the best least squares fit.

For example, one might want to rotate configurations for the same set of objects obtained from two different proximity matrices by an MDS procedure, or two factor solutions based on the same unit.

PROCROT will echo input, print the rotated matrix, plot it if desired, and provide a number of indices of goodness of fit. It is a portion of a larger program written by Professor Norman Cliff of the Psychology Department at the University of Southern California.

Time and Region: To rotate one 40 x 3 configuration to another 40 x 3 configuration, TIME = 0.70 CPU seconds, REGION = 76K.

Documentation: Users' instructions are contained as comments in the source. For a discussion of theory, see N. Cliff, Psychometrika, Vol. 31, 1966, pp. 33-42.

PROCROT is extremely easy to use. Note, however, that if configurations are to be read from cards or disc, the present version assumes that the input format accompanies the configuration (because it was modified to read POLYCON output, which automatically writes the format on the punch dataset). If the configurations are bodily included with the run, the input format must be specified just ahead of the configuration to be read.

PROCROT was written by N. Cliff and modified by C. Frost.

DISTPROG

DISTPROG is a program written to compute distances between pairs of objects, based on specified variables, for input to MDS programs; i.e., it computes derived distances.

It will compute distances from variables in raw form or scaled by their standard deviations (to compensate for irrelevant differences in scale). It will also, if desired, compute distances between variables from the same or another dataset, scaled or not, and form them into a supermatrix.

Output, in addition to the distance matrix, includes echoes of input and output formats, standard deviations of variables, and echo of instructions.

Time and Region: No details have been saved, but to compute a 40 x 40 distance matrix over ten variables took much less than 1 machine unit in prime time.

Documentation: Users' instructions appear at the beginning of the source listing.

DISTPROG was written by C. Frost.

TRANSPOS

TRANSPOS is basically an auxiliary program to set up runs for PROFIT (q.v.), although it can be used simply to transpose a matrix.

PROFIT requires as input a configuration matrix which is N by K, K dimensions for N units; and a property matrix, P by N, for P properties or scales. Since data are commonly stored and used N by P, where P is greater than the number of scales to be fit to a configuration, it is usually necessary to select a submatrix from the full N by P, and to transpose that submatrix.

Basically, then, TRANSPOS is intended to select the submatrix and transpose it. However, it is designed to do more: It will take as direct input a full set of PROFIT control cards, select and transpose the submatrix, read the configuration, and assemble and write out the entire PROFIT run on a dataset called PROFIT.RUN 1, which can then be run simply by executing the actual PROFIT.RUN job which contains the JCL, jobcard and SYSIN defined as PROFIT.RUN 1.

No details are available, since TRANSPOS output is hardly ever printed, but it runs in LT 100 and uses LT 1 machine unit.

Documentation: Users' instructions appear as comments at the beginning of the source listing.

Note that, at present, the output dataset is 10, the property input is 09, and the configuration input is 11; these may be changed as shown in the users' instructions.

TRANSPOS was written by C. Frost.

BIBLIOGRAPHY

Anderson, S., "Assessment for Personal and Educational Development," CIRCUS: Comprehensive Assessment in Nursery School and Kindergarten, proceedings of a symposium presented at the American Psychological Association Convention, Montreal, Quebec, August 21, 1973.

Anderson, S., and S. Messick, "Social Competency in Young Children," Developmental Psychology, Vol. 10, No. 2, 1974, pp. 282-293.

Anderson, G. H., and H. J. Walbert, "Learning Environments," in H. J. Walberg (ed.), Evaluating Educational Performance, McCutchan Publishing Corp., Berkeley, California, 1974.

Bargen, M., and H. J. Walberg, "School Performance," in H. J. Walberg (ed.), Evaluating Educational Performance, McCutchan Publishing Corp., Berkeley, California, 1974.

Barker, P., S. Abramowitz, T. K. Bikson, R. L. Crain, C. Frost, and S. Pelavin, "Concerning Powerful Nonmetric Methods of Analysis," The Rand Corporation, unpublished paper, May 1975.

Barker, P., and T. K. Bikson, Measuring Classroom Diversity and Individualization: The LUMPINESS Index, The Rand Corporation (in preparation).

Barker, P., T. K. Bikson, and J. Kimbrough, Training Manual for the Alum Rock Classroom Observation Study, The Rand Corporation, P-6183 (forthcoming).

Bass, G. V., A Study of Alternatives in American Education, Vol. 1: District Policies and the Implementation of Change, The Rand Corporation, R-2170/1-NIE, April 1978.

Bennett, S. N., and J. Jordan, "A Typology of Teaching Styles in Primary Schools," British Journal of Educational Psychology, Vol. 45, 1975, pp. 20-28.

Berman, P., M. W. McLaughlin, G. V. Bass, E. Pauly, and G. Zellman, Federal Programs Supporting Educational Change, Vol. VII: Factors Affecting Implementation and Continuation, The Rand Corporation, R-1589/7-HEW, April 1977.

Bijou, S. W., R. F. Peterson, and M. H. Ault, "A Method to Integrate Descriptive and Experimental Field Studies at the Level of Data and Empirical Concepts," Journal of Applied Behavior Analysis, Vol. 1, No. 2, 1968, pp. 175-191.

Brandt, R. M., "Observational Methodology for Evaluation of Early Childhood Programs," Journal of Research and Development in Education, Vol. 6, No. 3, Spring 1973, pp. 94-109.

Bredo, E., Collaborative Relationships on Teaching Teams: Implications for Collegial Influence, Team Morale, and Instructional Practices, Technical Report No. 45, Stanford Center of Research and Development in Teaching, Stanford, California, 1975.

Brophy, J., "Achievement Correlates," in H. J. Walberg (ed.), Evaluating Educational Performance, McCutchan Publishing Corp., Berkeley, California, 1974.

Brophy, J. C., et al., "Classroom Observation Scales: Stability Across Time and Context and Relationships with Student Learning Gains," Journal of Educational Psychology, Vol. 67, No. 6, 1975, pp. 873-881.

Calfee, R. C., and K. A. Hoover, RAMOS Event Form, Stanford University Press, Stanford, California, 1974.

Census of Population and Housing, San Jose SMSA, 1970, Department of Commerce, Bureau of the Census, P-HC (1-190), Washington, D.C., April 1972.

Cochran, W. G., Notes on Least Squares Linear Regression and Analysis of Variance, Harvard University, Department of Statistics, Cambridge, Massachusetts, 1970.

Cohen, D., "Social Experiments with Schools: What Has Been Learned?" in A. M. Rivlin and T. M. Timpane (eds.), Planned Variation in Education: Should We Give Up or Try Harder? The Brookings Institution, Washington, D.C., 1973.

Cohen, S. R., "Individualizing Instruction: Do Aides Make a Difference?" American Educational Research Association, 1973.

Condry, J., "Enemies of Exploration: Self-Initiated Versus Other-Initiated Learning," Journal of Personality and Social Psychology, Vol. 35, No. 7, 1977, pp. 459-477.

Draper, N. R., and H. Smith, Applied Regression Analysis, John Wiley and Sons, Inc., New York, 1966.

Fiedler, M. L., "Bidirectionality of Influence in Classroom Interaction," Journal of Educational Psychology, Vol. 67, 1975, pp. 735-744.

Fine, H. J., and C. N. Zimet, "A Qualitative Method of Scaling Communication and Interaction Process," Journal of Clinical Psychology, Vol. 12, 1956, pp. 268-271.

Flanders, N., Analyzing Teacher Behavior, Addison-Wesley, Reading, Massachusetts, 1970.

Foa, U. G., "Convergences in the Analysis of the Structure of Interpersonal Behavior," Psychological Review, Vol. 68, 1961, pp. 341-353.

Glass, G., "Teacher Effectiveness," in H. J. Walberg (ed.), Evaluating Educational Performance, McCutchan Publishing Corp., Berkeley, California, 1974.

Heller, M. S., and M. A. White, "Rates of Teacher Verbal Approval and Disapproval to Higher and Lower Ability Classes," Journal of Educational Psychology, Vol. 67, 1975, pp. 796-800.

Johnson, D. W., et al., "Effects of Cooperative Versus Individualized Instruction on Student Prosocial Behavior, Attitudes Toward Learning, and Achievement," Journal of Educational Psychology, Vol. 68, No. 4, 1976, pp. 446-452.

Joreskog, K. G., G. T. Gruvaeus, and M. VanThillo, ACOVS: A General Computer Program for the Analysis of Covariance Structures, Educational Testing Service, ETS Research Bulletin 70-15, Princeton, New Jersey, 1970.

Krantz, P., and T. Risley, "The Organization of Group Care Environments: Behavioral Ecology in the Classroom," American Psychological Association, 1972.

Longabaugh, R., "A Category System for Coding Interpersonal Behavior as Social Exchange," Sociometry, Vol. 26, 1963, pp. 319-344.

Mason, W. S., "Problems of Measurement and the NIE Program," NIE, August 29, 1973 (mimeograph).

Moustakas, C. E., I. E. Siegel, and H. D. Schalock, "An Objective Method for the Measurement and Analysis of Child-Adult Interaction," Child Development, Vol. 27, No. 2, 1956, pp. 108-135.

Prescott, E., et al., Assessment of Child-Rearing Environments: An Ecological Approach, Pacific Oaks College, Pasadena, California, 1975.

Prescott, E., et al., Group Day Care as a Child-Rearing Environment, report prepared for the Children's Bureau, Social Security Administration, Pacific Oaks College, Pasadena, California, 1967.

Rasmussen, R. L., S. Abramowitz, and E. Levinson, "Organization, Management, and Incentives in the Alum Rock Schools," The Rand Corporation, unpublished paper, May 1977.

Risley, R., and M. F. Cataldo, Planned Activity Check, Center for Applied Behavior, Kansas City, Kansas, 1973.

Rosenshine, B., and N. Furst, "The Use of Direct Observation to Study Teaching," in M. W. Robert Travers (ed.), Second Handbook of Research on Teaching, Rand McNally, Chicago, Illinois, 1973

Rubin, D. B., "Estimating Causal Effects in Randomized and Non-Randomized Studies," Journal of Educational Psychology, Vol. 66, 1974, pp. 688-701.

Sequoia Institute and Voucher Demonstration Project, Voucher Choices: A Family Directory of Mini-Schools, Department of Research and Evaluation, Alum Rock Union Elementary School District, San Jose, California, 1975.

Shepard, R. N., "Introduction to Volume I," in R. N. Shepard, A. K. Romney, and S. B. Nerlove (eds.), Multidimension Scaling: Theory and Applications in the Behavioral Sciences, Vol. I, Seminar Press, New York, 1972.

Siegel, L., and L. C. Siegel, "A Multivariate Paradigm for Educational Research," Psychological Bulletin, Vol. 68, No. 5, 1967, pp. 306-326.

Stallings, J., and D. Kaskowitz, Follow Through Classroom Observation Evaluation, 1972-1973, prepared for the Division of Elementary and Secondary Programs, Office of Planning, U.S. Office of Education, Stanford Research Institute, Menlo Park, California, 1974.

U.S. Commission on Civil Rights, Teachers and Students, Report V: Mexican American Education Study, Washington, D.C., 1973.

Waxler, N. E., and E. G. Mishler, "Scoring and Reliability Problems in Interaction Process Analysis: A Methodological Note," Sociometry, 1965, pp. 28-40.

Weiler, D., et al., A Public School Voucher Demonstration: The First Year at Alum Rock, The Rand Corporation, R-1495-NIE, June 1974.

White, S., et al., Federal Programs for Young Children, prepared for the U.S. Department of Health, Education, and Welfare, The Huron Institute, Cambridge, Massachusetts, 1973.

Young, F. W., Conjoint Scaling, University of North Carolina, the L. L. Thurstone Psychometric Laboratory, Chapel Hill, North Carolina, 1973.